Customizing News Content for Individuals and Communities

Marko Turpeinen

Helsinki University of Technology Faculty of Information Technology Department of Computer Science

Dissertation for the Degree of Doctor of Technology to be presented with due permission for public examination and debate in Auditorium T2 at Helsinki University of Technology, Espoo, Finland, on the 25th of February, 2000.

Turpeinen, M. Customizing News Content for Individuals and Communities. Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 103., Espoo, 2000. Published by the Finnish Academy of Technology. ISBN 951-666-531-4, ISSN 1456-9418.

Keywords: customized electronic publishing, metadata, news augmentation, customer modeling, community of practice

Abstract

As media organizations embrace the Internet as a viable media, new models of publishing are being investigated. This is due both to the need to find a new economics of customerdriven publishing services and to the emergence of a more demanding and technologically empowered customer. Traditional media channels have not been suitable to efficient customization. News content has been produced for global or national audiences, or smaller geographic communities, like local newspapers. Internet-based customized news services can serve the special needs of individual news customers, as well as small, and often geographically dispersed, communities. Customized forms of information selection and presentation increase the perceived relevance of media content, and provide new and potentially more efficient tools for learning and creation of new knowledge.

News content can be customized by focusing customer's attention to a selected subset of news stories, or by augmenting selected news with related material or illustrative comparisons. Focusing can include filtering, which means showing only the stories that are assumed to be most relevant to the customer, or prioritization, meaning highlighting and emphasizing these stories. The goal of customized news augmentation is to relate the news stories better to what the user already knows. Augmentation combines continuous news streams with automatically selected personal or communal contextual information from heterogeneous sources. Financial news augmentation is an example of a service that combines real-time business news with customized background information, term explanations, and comparisons about numerical data and organizations appearing in the news.

For the purposes of customized services, rich and dynamically changing user and community models are gathered about the media customers. User models can consist of special interests, expertise level, previous activity, and community context of any individual user. Community models include information about function, identity, discourse, and shared values of the community. When the news content is matched with customer profiles, a rich description is needed. This calls for deep multi-dimensional metadata to be used as a basis for customization.

News providers have multiple options to position themselves in customized media. Internet is driving towards a decentralized and community-oriented mode of publishing, and news organizations are losing their traditional position as the sole gatekeeper of information. Rich customer profiles and high quality metadata about the media content become increasingly valuable, when news organizations are facing new competition as new intermediaries are collecting this necessary information for customization. The media companies have an advantage, if they can innovatively combine the skills of professional journalists with customized services that meet the special needs of individuals and small communities.

Acknowledgements

The world of electronic media has radically changed during the past five years that I have spent working on this dissertation. The unprecedented surge of interest and rapid deployment of Internetbased communications has been (and continue to be) astonishing. I joined the Ota Online project at the Helsinki University of Technology (HUT) in the beginning of 1994. The futuristic ideas of customized information services of that time are now almost commonplace and widely available on the Internet. This rapid pace of change and innovation has been a challenge from the researcher's point of view. On the other hand, I cannot think of a better way to participate in a communications revolution.

I consider myself extremely fortunate to have been able to work with such a large number of knowledgeable and passionate people. First of all, I would like to thank Professor Reijo "Shosta" Sulonen for keeping me on the right path (and almost in schedule) with his guidance, encouragement, and wisdom. I also want to express my gratitude to my pre-examiners, Professor Nils Enlund, who introduced me to the world of electronic publishing in 1993, and Professor Tapio Varis, for their constructive comments and suggestions for improvement on the thesis manuscript.

My former colleagues at the Ota Online project, Mari Korkea-aho, Tuomas Puskala, and Janne Saarela, provided an inspirational working environment when the Web was still very young. More recently, the members of the SmartPush team at HUT, Sami Jokela, Teppo Kurki, and Eerika Savia, have been valuable research collaborators and commentators of my work. Many thanks also to Ari Heinonen from University of Tampere for his thoughtful and clarifying remarks.

I want to express my gratitude to Matti Packalén and Jaakko Nieminen for providing a working environment for innovative thinking and intellectual development at Alma Media Corporation. Special warm thanks to Jaakko Nieminen, for your constant encouragement, and for giving me the space and the time that I needed to make this possible. Many of my colleagues at Alma Media have also directly contributed to this work. I would like to thank Marja Heinonen, for enlightening discussions on the difference between engineers and journalists, and for her constructive feedback on my work. Mauri Mattsson, JR Leskinen, Hannu Olkinuora, and Eero Tuomisto have been invaluable sources for ideas and inspiration since the beginning of the Ota Online project. Hannele Vihermaa and Markku Ylinen have taught me a lot about the nature of news and information. I also wish to thank Timo Saari, the other half of "the research twins" at Alma Media, for intellectual stimulation and for shaping my thinking in a fruitful dialogue.

I feel very privileged to have studied and worked at the MIT Media Laboratory. I would like to thank Walter Bender, Guillaume Boissière, Pascal Chesnais, Jack Driscoll, Ingeborg Endter, Dan Gruhl, Ken Haase, Martin Hadis, Pattie Maes, Warren Sack, and many others at the lab for an endless source of fresh ideas and for your helpful feedback. I am grateful to all of you for making

the Media Lab such a unique and creative environment for working, learning, and occasional softball.

I want to thank my parents, my brother, and my grandfather for their support and caring. My children, Aleksi and Laura, have constantly reminded their dad about the important things in life. And finally, my heartfelt thanks to Hanna - without your encouragement and patience this work simply would not have been possible.

Cambridge, January 25, 2000.

Marko Turpeinen

Contents

Part I

1 Introduction				
1.1	Definitions for the study of news customization	3		
1	.1.1 Content	4		
1	.1.2 Presentation	4		
1	.1.3 Delivery	5		
1	.1.4 Can content be separated from presentation and delivery?	5		
1.2	Research objectives and method	6		
1.3	Thesis overview	7		
1.4	Key contributions of this thesis	8		
2 E	Examples of news filtering and news augmentation	10		
2.1	Filtered news	10		
2	Adaptive news filtering	13		
2.2	Augmented news	14		
2	2.2.1 Customized news contextualization	17		
2.3	Conclusions on the example services	19		
3 L	Understanding news content	21		
3.1	News content	22		
3.2	Uses for news	25		
3.3	Community-centric news	26		
3.4	News contexts	27		
3.5	Decontextualized news	32		
3.6	Summary of findings	33		
4 L	Learning and knowledge creation through news	35		
4.1	Schemata, stories, and scripts	36		
4.2	Conceptual models for sensemaking	38		
4	.2.1 Ontologies	20		

4	.2.2	Sensemaking	41			
4.3	Soc	ially situated knowledge and expertise	42			
4	.3.1	Communities of practice	43			
4	.3.2	Boundaries and membership	44			
4	.3.3	Situated expertise and knowledge creation	46			
4.4	Imj	plications for customized news	48			
5 0	Custom	izing news				
5.1	Foc	using for individuals	50			
5	.1.1	Filtering	50			
5	.1.2	Prioritizing	53			
5.2	Foc	using for communities	54			
5.3	Au	gmentation for individuals	56			
5.4	Aug	gmentation for communities	59			
5.5	Sur	nmary of findings	60			
6 I	mplem	enting news customization	62			
6.1	Cus	stomer modeling	62			
6	5.1.1	User modeling	63			
6	5.1.2	Community modeling	69			
6.2	Сог	nbining individual and community profiles	74			
6.3	Сог	ntent modeling	76			
6	5.3.1	Ontologies for semantic metadata	76			
6	5.3.2	Adaptive and dynamically created hypermedia news	78			
6	5.3.3	Matching customer profiles and content profiles	78			
6.4	Exa	mple systems				
6	5.4.1	Adaptive news filtering	79			
6	5.4.2	News augmentation system	79			
6.5	Sur	nmary and implications	85			
7 N	Aodels	for publishing customized content				
7.1	Bro	padcasting model	86			
7.2	Individualized model87					
7.3	Inf	omediary model	88			

7.4 Community-or	riented model	
7.5 Community-m	nediator model	90
7.6 Community-co	entric model	91
7.7 Summary of fi	indings	
8 Conclusions		94
List of Figures		97
List of Tables	98	
References	99	

Part II

Summary of publications	102
J J J I J I I I I I I I I I I I I I I I	

Publication 1.

Turpeinen, M., Saarela, J., Korkea-aho, M., Puskala, T., Sulonen, R. (1996). *Architecture for Agentmediated Personalised News Services*, PAAM'96: The First International Conference on Practical Application of Intelligent Agents and Multi-Agent Technology, 22-24 April 1996, London, UK.

Publication 2.

Saarela, J., Turpeinen, M., Korkea-aho, M., Puskala, T., Sulonen, R. (1997). *Logical Structure of a Hypermedia Newspaper*, Information Processing and Management. Vol. 33, No. 5, pp. 599-614.

Publication 3.

Kurki, T., Jokela S., Turpeinen M., Sulonen R. (1999). Agents in Delivering Personalized Content Based on Semantic Metadata, Intelligent Agents in Cyberspace, Papers from the AAAI Spring Symposium, Technical Report SS-99-03, AAAI Press

Publication 4.

Turpeinen, M. (1999). Augmenting Financial News for Individuals and Organizations, Special Issue on Intelligent Agents for Education and Training Systems, International Journal of Continuous Engineering Education and Life-Long Learning

Publication 5.

Jokela, S., Turpeinen, M., Sulonen, R. (2000). *Ontology Development for Flexible Content*, HICSS-33 Minitrack on Systems Support for Electronic Business on the Internet, January 4-7, 2000

Part I

1 Introduction

Early visionaries in the field of interactive and personalized computing, such as Vannevar Bush in *As We May Think* [Bush, 1948] and Douglas Engelbart in *Augmenting Human Intellect* [Engelbart, 1962], believed that human beings can be empowered and augmented by personal computational systems; tools, concepts, and methods that match individual's basic capabilities to their problems. Rapidly growing computer processing and storage capacity, the shrinking size of electronic components, and the universal connectivity of the Internet, have brought us close to realizing the ideas of Bush and Engelbart.

Computer-mediated networking between people has enforced existing physical communities [Engelbart, 1995; Bender et al., 1996; Schuler, 1994; Carroll and Rosson, 1996], and has helped in building entirely new virtual communities, meaning electronic arenas where people meet and participate in joint activities [Rheingold, 1993]. This is especially clear with the explosive growth of the World-Wide Web. These new forms of group communication and participation are fundamentally different than traditional mass media communications. Therefore, current media companies are struggling to understand how these communities operate, how media professionals could support their activities, and how the traditional media could benefit from this "grassroots activity".

The history of mass communication media is strongly influenced by technological breakthroughs: printing press, telegraph, radio and television broadcasting, and most recently packet-switched digital networks including the Internet. Traditionally new publications have been created based on the economic viability of the audience as either paying subscribers or as potential consumers of advertised products. Entry barriers of mass communication in advertising economics and limited frequency spectrum in broadcasting media have kept the number of media companies fairly low [Neuman, 1991]. Also, by their nature, the mass communication media have discouraged two-way communication and participation as the publishers have delivered their content to the consumers through one-way media channels.

Digital media and *Computer-Mediated Communication (CMC)* allow for unprecedented flexibility in media content creation and access. This flexibility of networked hypermedia is combined with a large potential for interactivity and customization. With widely used Internet-centered applications, it has become obvious that the computer is not just a task-oriented utility tool, but also a communication tool for a wide variety of needs, as Winograd [1996] writes:

The computer (with its attendant peripherals and networks) is a machine for communicating all kinds of information in all kinds of media, with layers of structuring and interaction that could not be provided by traditional print, graphic, or broadcast media.

Tetsuro Tomita of Japan's Ministry of Posts and Telecommunication proposed in 1980 the diagram depicted in Figure 1 [Tomita, 1980]. The horizontal axis organizes media by the size of the audience, and the vertical dimension shows the delay in seconds between the event in real life and its appearance as a media product. Most media products have been geared towards large audiences, the only exception being small circulation printed newsletters and periodicals, which typically are either non-profit ventures or very targeted business or scientific publications. This *media gap* identifies an area in which the needs of relatively fast-paced small-group and special-interest-group communications are not met by traditional media. Tomita proposed that computer mediated electronic communications will fill the gap, as is clearly happening with Internet-based contents. Usenet newsgroups and web sites¹ exist for professions, hobbies and interests of all sorts. Most of these forums are open to public around the globe. Many of them are oriented towards, or run by a *community of practice*, which means a group of people with *joint enterprise or purpose*. However, it is rare that these community efforts to fill the media gap are run by, or serviced by, media companies, and the potential for co-beneficial relationship between traditional mass media content producers and small and active community publishers has not been realized.



Figure 1. The media gap, adapted from [Neuman, 1991]

As media organizations increasingly consider the Internet as the publishing platform of the future, new models of electronic publishing are being investigated. New media technologies, such as Internet, can

¹ The terms *World-Wide Web*, *WWW*, and *the Web* are synonymous.

result in a decline of the importance of mass communications media. For example, the synchronous media consumption patterns of everyone watching the same news broadcast at a given hour have already broken down as the number of news broadcasting channels has exploded. This trend towards more fragmented and asynchronous media usage will be enforced with the increased possibilities provided by the Internet.

The intermediate level of group participation and community communications is between personal and mass media. One promising option is to exploit the dynamic nature of network media and to customize the media experience to individuals and communities. This *mass customization* [Pine, 1992] of media content and services has been impossible with traditional media technologies, whereas computational media can know about its users and can tailor the message so that the news are made more relevant and more understandable [Bender, 1996].

Customized information products can partially fill the media gap, as they meet different requirements of individuals or communities. Customized information services move away from broadcasting-oriented information production, as the producer tailors information according to profiles of customers and customer groups. The opportunities in providing new kinds of adaptive media products, which empower the individual customers and strengthen community communications, are broad.

Over the history of communications, *news reporting* has evolved into one of the main information products offered by the media industry. News producers range from small community-oriented local newspapers to worldwide news organizations producing news 24 hours a day. The news industry has become an important actor in the society, mainly because the news producers have the power to choose what topics and events are considered important (i.e. *newsworthy*).

1.1 Definitions for the study of news customization

This study concerns *customized news services* that are tailored according to *customer profiles*. The concept of *news* defies a simple definition and is discussed in detail in chapter 3. A *news customer* is an individual or a community (consisting of members who are individuals) that is an end user for news services. A *news producer* is an individual or a community that provides content for news customers. The news producer can be a professional organization, such as a media company, networked community, or an individual looking for an audience for their news reporting². A *news intermediary* is an entity providing a service that integrates news information from multiple news producers, and packages and distributes the content but does not necessarily produce news. The customized news services can also be *adaptive*, i.e. they change behavior according to observations about the customers. Although I concentrate on news content, the principles discussed in this work can be applied to a wide variety of network-based customized communications, such as electronic commerce and customized advertising.

² Matt Drudge's Web site <u>http://www.drudgereport.com</u> is a well-known and controversial example.

In principle, there are three potential approaches to make a tailored news service:

- content can be selected, grouped, and organized according to customer profile,
- presentation can be tailored to suit the needs and preferences of the customer, and
- *delivery* methods can be tailored by media platform capabilities, update time and frequency, and cost.

1.1.1 Content

This study focuses on news content from two perspectives: *focusing* and *augmenting*. News filtering is an example of a focusing service, where a subset of the full news content flow is automatically selected and presented to the customer (Figure 2). The goals of focusing are to ease the effects of information overload, and to serve the customer with specific pre-determined information needs.



Figure 2. The role of focusing and augmenting in news services

In a *news augmentation* service, the news content is explained and contextualized with related information [Elo, 1996]. This information is selected and presented in a customized manner according to the customer profile. The *currently focused story* is enriched with additional information from heterogeneous data sources. These augmentations are presented in a manner that puts the story in context, and make it more understandable for the customer.

1.1.2 Presentation

Customization of the presentation means packaging the news contents into a form which best suits the customer needs and the context of use. There are many needs for tailored presentation, such as

• mechanisms for providing equal access to the content for disabled users; *Web Access Initiative (WAI)* addresses this problem of making the Web accessible to people with disabilities using guidelines for

creating documents that can be transformed according to user needs [World Wide Web Consortium, 1999a], and

• personal preferences for layouts and color schemes; *Cascading Style Sheets (CSS)* specification allows personalized presentation styles of document contents [World Wide Web Consortium, 1999b].

As an example of how the layout and style of news presentation can be tailored according to customer's personal taste, Weitzman and Wittenburg [1995] presented a system that automatically generates different spatial layouts and graphical styles for a multimedia document. This provides different results based on display device, browsing environment, or publication style. For example, the same story can be presented in the styles of Scientific American or Wired magazine.

1.1.3 Delivery

Customer can prefer different devices for different content. The motivation for this can result from media platform capability limitations, such as bandwidth, or personal information consumption habits. Information with high data rate (video, high-resolution images) is replaced by less data intensive equivalents (audio and text script, low-resolution images) for users with low-bandwidth access to network. Customers may want only news headlines on a low-bandwidth mobile device, while others are willing to wait longer for the transmission delay and access the full contents on the mobile device. Time and frequency for content updates are also valid parameters for delivery personalization. As an example, *Wireless Application Protocol Forum* $(WAP)^3$ has defined various standards such as the Wireless Markup Language (WML) and Wireless Application Environment (WAE) to deal with specific needs of low-end devices and low-bandwidth network connections, especially in mobile environments.

1.1.4 Can content be separated from presentation and delivery?

From the technical perspective, these three aspects – content, presentation, delivery – can be considered different. However, it can easily be argued that the media experience consists of the combination of all three, and that for many practical applications they are inseparable. This is the case with the newspaper. The content and the presentation are entangled, as there are plenty of subtleties in newspaper page formatting and layout design. The readers have accustomed to a specific type of connotations that the newspaper format has as a communication media.

SGML, and more recently XML, separates the logical document elements from presentation rules, but allows sophisticated mechanisms through style sheets for providing the expressiveness familiar to print products onto digital media [Goldfarb and Prescod, 1998]. Structured documents (XML) and style sheets

³ WAP Forum can be found at <u>http://www.wapforum.org/</u>

(CSS) provide a solid basis for customization. However, not all subtle meanings preserved in layout, colors, imagery, can be encoded as logical XML structures to be rendered with a style sheet.

As an example of a customized electronic news service that combines presentation and customized content, Södergård et al. [1999] use a layout engine for personalized on-screen pagination, using a constraint-based approach for constructing multimodal presentations.

1.2 Research objectives and method

This thesis is a combination of research related to news customization that I have done in 1995-1999 at Ota Online and SmartPush projects at Helsinki University of Technology, Daily Catch and Silver Stringers projects at MIT Media Laboratory, and internal projects at Alma Media Corporation.

This work explores the options available for building customized news services. In addition to challenges of technical implementation, this thesis builds upon prior research in a number of fields, mainly cognitive theories of learning and mass communications studies. News customization requires understanding of existing media, as well as human information processing in a social setting. This is motivated by research findings that broadcast and print news content is sub-optimal from the perspectives of learning, remembering, and knowledge creation. New media forms, such as the Web, allow news reporting that can be more engaging and informative than the news provided by newspapers and broadcasting.

First task is to build a hypothesis of potential benefits for news customers that might be achievable through news customization. This hypothesis is based on the analysis of main characteristics of news genre, and on cognitive theories of story-based understanding and situated knowledge creation in social environments. This social aspect of news is considered through communities that need to share information, especially due to shared interests and practices.

This work considers the added customer value of customization in new kinds of news applications, and what is their impact on news producers. This is done mostly using examples of research prototypes, as well as first generation commercial products and services.

After making explicit the map of content customization, and main challenges involved in implementing customized news, I evaluate the impact of these services to traditional media companies. The rise of the Web, personalization, and active and participating communities that are creating their own content, can all be seen as a threat or as an opportunity for traditional media. My hypothesis is that customization is a key to value-added publishing in the digital media. It is crucial for media companies to understand the interplay of individual and community needs for customization. I present a synthesis of available options for content producers in building customized news services, including an overview of the possibilities and limitations of technical mechanisms for dealing with the complex dynamic nature of customer profiling. This analysis can be used as a basis for recommendations for news producers in delivering customized information.

The nature of this work is explorative and descriptive. This thesis does not include a systematic study of the effects of implemented and analyzed systems, especially from the perspective of learning from news, although some of the implemented prototype systems are explicitly addressing the problem of learning. This empirical evaluation from the perspective of learning was outside the scope of this study. If empirical results of the efficiency of other example news systems are available from work done by other researchers, they are presented in this thesis.

The technological contribution of this work is to introduce implemented prototype systems, and to discuss the major challenges involved in building them. The areas of computer science relevant in the framework of this thesis are mostly discussed in accompanied publications, as listed below:

- agent technology for information filtering and information augmentation (*Part II: Publications 1, 3, and 4*),
- representation of content metadata and content ontology (Part II: Publications 2, 3, and 5),
- adaptive customer modeling using machine learning (Part II: Publications 1, 3, and 4), and
- *dynamic hypermedia* for customized news presentations (*Part II: Publications 2 and 4*).

1.3 Thesis overview

This thesis is divided in two parts. Part I presents an introduction to the area of customized news content. In chapter 2, I introduce two state-of-the-art news services on the Internet (as of Spring 1999) by discussing their strengths and shortcomings. The role of adaptivity and information augmentation in customer-oriented news media is introduced in relation to the example services presented.

Chapter 3 is a brief introduction to the news media genre, which sets the ground for the discussion on customized news content in chapter 5. News genre is analyzed from the points of view of customer needs for news, news content and contexts, the narrative nature of news media, and news as a potential resource for knowledge creation. Finally, I introduce one of the most important trends of networked communication, which is the role of active communities that are not only consuming, but also producing, news and other kinds of information.

Chapter 4 introduces theories of situated learning and communities of practice, and discusses their relation to news media.

Chapters 5 and 6 are the core of this thesis. New forms of communications allowed by digital two-way networks, such as the Internet, change the possibilities for content creation and the structure of media industry. Mechanism of simultaneous customization for individuals and communities is introduced. The most relevant questions for these chapters are: How do we find most relevant news articles for each individual user based on what we know about the contents and what we know about the customers? How can the news be augmented in a customized manner? What is the relationship between user profiles and

community profiles? How much weight should be given to the community context of the user instead of the personal context? What is the relationship between the two?

Chapter 6 introduces the software architecture for building customized augmentations. The role of metadata-based information integration is essential. Software agents are used as building blocks for these customized services, as profile-based news services are implemented as a set of interoperating agents. A financial news augmentation system prototype for Kauppalehti Online is also discussed in chapter 6, and in more detail in *Part II: Publication 4*.

The recommendations for publishers are presented in chapter 7. This chapter also discusses the overall implications of customization for news media. The conclusions from this study are found in chapter 8.

Part II of this thesis contains the individual publications that present results from different technical research areas of the thesis. Part II begins with a brief description of each publication and about the author's contribution to them. Then, each publication is presented in full.

In addition to traditional references, URLs are included in the references and as footnotes where applicable. All URLs have been checked for validity in November 1999.

1.4 Key contributions of this thesis

To finish this introduction, I have summarized the main themes and the related findings of this work, which will be motivated and justified in the chapters to follow.

Customization strengthens customer relationship for electronic publishing

Internet challenges traditional methods of publishing. Publishers are looking for new ways to attract and keep their customers in the electronic world. Customization increases the value of information products, and can be targeted towards individual customers and customer groups. News producers have not traditionally considered their product as something that can be customized. Because customization is always dependent on acquisition and creation of customer profiles, which are built over time, it becomes a critical function for content providers in keeping their customers and in differentiating themselves from their competitors.

Rich semantic metadata combined with editorial skills are required for high-quality customized news

Descriptive data about media resources is needed to create high quality customized news services. Metadata can be created automatically or semi-automatically. However, quality assurance of metadata and ontology creation that is necessary for high quality metadata requires professional human expertise. This is a big organizational change facing news producers, if they are to provide attractive customized services.

Creative ways in combining the skills of professional journalists and automated software tools for metadata creation and content customization is the key to success.

Personalization should be combined with community-oriented customization

Customized news service can be used to identify, better serve and strengthen different kinds of communities. Active, participating, and creative individuals and communities are challenging the traditional models of publishing and journalism. Hybrid customization, which combines individual profiles and community profiles, takes into account the dual nature of news – it is both individual and social.

Customized augmentation is a natural way to enhance news content

There is more to content customization than filtering or recommending information, although information filtering does fill an important subset of customer's need for news. Customization mechanisms for focused news delivery should be combined with news augmentation that expand and contextualize the news in order to make them more understandable to the readers. Augmentations relate the news story to the individual and community context of the receiver. This imitates the natural mechanisms of human storytelling.

2 Examples of news filtering and news augmentation

Two fairly sophisticated commercial news services for the Web are discussed in this chapter for comparison with the work presented in this thesis. They illustrate how far the news industry has embraced the Web as a new medium for expression and tailored content, and show how personalized news filtering and news contextualization can be implemented as web services.

2.1 Filtered news

There are multiple services on the Web for personalized news filtering. These filtering services are typically based on explicitly defined individual's interest profile. As an example, I discuss the popular *My Yahoo!* service of Yahoo!, as it was in April 1999. The service consists mainly of news and other topical interests that the customer can subscribe to. Figure 3 shows an example of My Yahoo! front page tailored for my preferences. Customer makes selections from a pre-determined content category. News and other information items are matched against these selections. Matching is based on taxonomy metadata that Yahoo! or the content producer organization has created for the content.



1. Personalization options can be set for content and layout.

- 2. Weather information from specified locations in degrees Celsius.
- 3. Latest sports scores from selected teams.
- 4. News headlines sorted according to category and source.
- 5. Personalized ad based on the geographic location of the reader. Selected ad is for Yahoo! Sweden since there is no ads for customers from Finland.

Figure 3. The front page of My Yahoo!

The personalization options are in the dimensions of content and presentation. My Yahoo! is accessible only in HTML format through a Web browser, so it is not tailored for different types of content delivery. The role of customized delivery presentation becomes more important as the services are going to be used from multiple media platforms with varying end device and bandwidth capabilities.

The content categories in My Yahoo! are fairly broad, but they lack depth since there is no hierarchical structure for concepts. At first glance this is somewhat surprising, since the core strength of Yahoo! has been the capability to categorize Web sites according their proprietary ontology of content. The reason for this shallow categorization is that Yahoo! acts as an information broker between the content producers and the customers. Since Yahoo! doesn't produce news content, they rely on the categorization that is done at the news producer's location.



Figure 4. Selecting the contents for the front page of My Yahoo!

The selection of content is made from a list of possible content blocks (Figure 4). The sources for contents are mostly not mentioned in this list. Selections for content blocks can be on or off. The news content block can have multiple topic areas that are either source-specific (like "Business from Reuters") or combined from different sources (like "Automotive"). The user can also select how many headlines are shown for each topical interest area in the news section of the My Yahoo! front page (Figure 5).



Figure 5. Selecting news by topic areas for personal front page



Figure 6. Customizing the layout of the front page of My Yahoo!

The customer can re-organize the display order of content blocks (Figure 6). The page layout consists of two columns and the arrangement of content blocks in these columns can be changed, as well as some minor options - like colors and greeting messages.

The customer can re-organize the display order of content blocks. The page layout consists of two columns and the arrangement of content blocks in these columns can be changed, as well as some minor options - like color settings for text and page background, as well as the greeting message displayed.

2.1.1 Adaptive news filtering

Overall, My Yahoo! is a good example of personal information brokering service that concentrates mostly on content, with some parameters for layout personalization. User interfaces for news services are becoming dynamic, and can be finely tailored to a user's wants and needs. My Yahoo! allows the user to customize their view of the site. This is achieved using standard components that are glued together into a personalized interface. My Yahoo! lets users specify their interests so that they would be notified by types of news, stock exchanges, sports, weather and other types of information they are interested in.

My Yahoo! is a good service as it stands. It provides a concise package of news content from the personal interest areas of the customer. However, there is room for improvement. The main shortcomings of My Yahoo! are in the following areas:

- *Adaptivity*. My Yahoo! is not adaptive. There is no dynamic learning or adjusting of interest profiles over time. All changes to the profile are done manually in the profile maintenance section. Adaptive news filtering system can register, analyze and classify the behavior of the user and update the user model accordingly.
- *Customizable ontologies*. My Yahoo! uses a simple pre-determined taxonomy of content areas. This fairly shallow category selection list is given to the reader in its entirety. The categories are also fully defined by Yahoo! and they cannot be complemented or mapped onto categories provided by others. Therefore, the categorization criteria cannot be changed or adjusted by the customer, and there is no easy way to combine interest profiles amongst different services.
- *Product management*. The system is assumed to stay the same over a long period of time. There is no clear mechanism for versioning the categories. In fact, the service provider gets easily stuck with the categories it started with.
- *Multiple usage contexts*. There is no possibility for multiple user roles. The profile at My Yahoo! stays the same no matter what the current context of the user is. Customer's needs might be very different if he/she is in a business role or in an entertainment-seeking role.
- *Depth for experts.* Information filtering is useful when the customer has well-specified needs, like local weather and sports scores of favorite teams. This service is especially valuable, when the rapid delivery of good quality punctual information is of high value for the customer. However, the topics

and the categorization in My Yahoo! lack detail to meet the needs of customers with more punctual information needs.

- Balanced serendipidity. The role of the service is to provide information on pre-determined areas of topical interest. However, filtering can restrict the worldview and the service can appear dull and uninteresting, exactly because it *is* what the reader has concentrated on before. Therefore, filtering systems typically meet only a subset of the need for news. Filtering should be combined with systems that explore and introduce new potentially interesting information domains to the user. Journalists are able to balance this need for serendipidity, to combine surprising and unexpected with known and familiar.
- *Privacy*. Many technical advances are made in allowing customers to use these services in anonymous or pseudonymous fashion. Personalization requires private information, and the users are increasingly more concerned about the possibility of misuse of their personal data, although My Yahoo! has a privacy policy that is posted on the Web site.

Many of these issues were addressed in Ota Online and SmartPush projects at Helsinki University of Technology, as reported in *Part II: Publications 1, 3, and 5*.

2.2 Augmented news

The World Wide Web has proven that hypertext presentations can be used as a powerful news media. The style of news reporting and storytelling that suits the capabilities of the World Wide Web has been formed. The combination of familiar features of newspapers, TV, radio, and magazines are put together in Web-based hypermedia news.

High-quality online news media provides both depth and breadth to the news content (further discussed in chapter 3.1). Skilled editors can combine a news story with background information from own sources of the news organization, as well as hyperlinks to carefully selected outside sources. Instead of conveying individual pieces of news, the journalists behind the news service offer contextualized and inter-linked explanatory material to the interested readers. The article becomes a gateway to exploring related information.

The following example from *CNN interactive* illustrates the state of hypermedia news in February 1999. The news story about a possible NATO air strike to Serbia contains a link to the "Conflict in Kosovo" theme at the top of the page with links to special related features. These theme augmentations are also shown as a blue sidebar further below (Figure 7). When the readers scroll down to the bottom of the page, they find that the story contains links to recent *related stories* from CNN archives. It also contains links to *related sites* elsewhere on the Web.

Top of the page

Bottom of the page



Augmentations in this CNN Kosovo story:

- 1. Link to the theme page "Conflict in Kosovo"
- 2. Direct links to the main features under the theme page
- 3. Hyperlinks to pages introducing military equipment planned to be used in the strike
- 4. Theme page links repeated as a sidebar
- 5. Related stories from CNN
- 6. Related sites outside CNN

Figure 7. Augmented news story from February 17, 1999 by CNN

The "Conflict in Kosovo" theme page is a good introduction to readers that are unfamiliar with the tragic events in Kosovo, or the Balkan crisis in general (Figure 8). It also serves as a useful reminder for more experienced readers in introducing the recent developments and the roots of the conflict.



- 1. Links to introduction of the conflict, "Key Players", commentary, and latest news
- 2. The military involvement and their arsenal in the Balkans area
- 3. Maps about Kosovo and its neighboring areas
- 4. Background and history of the conflict
- 5. Video clips further explaining the background of Kosovo conflict

Figure 8. "Conflict in Kosovo" theme page

The user can browse the map of Kosovo area, which contains descriptions of different geographical locations and their importance regarding the crisis. Also the historical context for the event is provided in the form of a timeline document that starts from the battle of Kosovo between Ottoman Turks and Serbs in 1389 and continues until 1999. The events are explained briefly, but the reader gets the sentiment of long-lasting and many-sided problem that is very difficult to solve (Figure 9).



Figure 9. An interactive map of Kosovo and surrounding areas, and a timeline of Kosovo history

Overall, the "Conflict in Kosovo" theme page significantly improves the understanding of otherwise isolated and brief news story. The approach is very similar to a magazine article that tries to uncover the underlying logic that has led to the current event. Multiple time scales are combined into a single presentation about the conflict.

Before the emergence of the Web, this would have required five different media:

- *television* and *radio* for broadcasted reporting and commentary,
- *newspapers* for in-depth storytelling,
- magazines for providing more context and high-quality photojournalism,
- encyclopedia to provide full depth to the event and explanatory material about basic concepts, and
- *computer terminal* to allow the user to search for related events from archives, and to participate in discussion about the events.

Now all these are combined into one digital media. The CNN contextualization of Kosovo events clearly fulfills the promise of hypermedia news reporting as described by Neuman [1991]:

What is truly unique to "new media", meaning the integrated network of digital communications technologies, is their capacity for integration and interconnection.

Considering the events that happened after this story was published (NATO started a bombing campaign against Serbia and the crisis escalated into a war), this contextualization is an example of good journalistic judgment by CNN on the importance of the events in Kosovo. The augmentations convey a sense of urgency and importance over other news themes, just like front page story placement, headline font size, or amount of air time given to a story: "*the story is so important, that you should understand what's behind all this*". Prioritization and contextualization are valuable journalistic services. However, the augmentations of this kind are rare, because they are expensive to produce.

2.2.1 Customized news contextualization

Automatic tools for contextualization can provide a similar kind of service to the reader, but will usually lack the journalistic selection and assessment of relative importance of this background information. However, I assume that this automatic contextualization is useful, if it is *available for all contents*, and if it can be *customized for each individual user*.

The Kosovo article from CNN provides a good example of the technical efficiency of the Web-based hypermedia news. Is there any room for improvement? Could the crisis be made even more understandable to the reader at different levels of context (Figure 10)?

News augmentations, as shown by CNN's Kosovo theme section, can help the reader to better understand the *event context* of the people that are affected by the news. What is the background for these new events? How does this new event relate to previous events? How could the media make the reader more aware of how people on different sides of the conflict feel? What would it be like to live in a city that is waiting for an air strike by NATO military forces? The Kosovo article provides background, but in a manner that is very typical to a media organization that is accustomed to news broadcasting model, as everyone receives the same contextualization.

The news happen in the event context that is observed and written about by the journalists. Journalists are typically "outsiders" that try to uncover what has happened and to tell in an understandable way a story about the event. The story is always a mediated version of what actually happened when it reaches the audience.

The audience for the news story consists of individuals and these individuals form different kinds of communities, each with its own context. Although CNN is a commercial news organization operating worldwide, the main target group for this special section are American readers and the style of reporting reflects the Western viewpoint on the crisis. The point of view given in Serbian news for the story was very different. As the situation in Kosovo aggravated, the Serbian media portrayed the Americans as the aggressors and eventually Americans systematically bombed Serbian media infrastructure. The *producer context* thus influences heavily the production methods and the style of reporting done by the journalists. How can a reader balance this producer context? Can combining and contrasting multiple sources do this?



Figure 10. Levels of context in news creation and reception

The Kosovo conflict can have many ties to reader's *community context*, and the reader can belong to many communities simultaneously. There can also be many reasons why the story is particularly interesting or controversial for a community. For example, there might be people living in reader's neighborhood that are from Kosovo, and from former Yugoslavia in general. What are they thinking? What are the similarities

and differences between the area of Kosovo and the reader's own neighborhood? Another example is the organization the reader is working for. What are the risks involved for our company regarding possible aggravation of this crisis? Have we done business with Serbia or it's neighboring countries recently?

CNN knows nothing about individual reader's personal background and interests, i.e. the *individual context*. It does not know how familiar the reader is with the subject, or how much the reader follows European news in general. The reader might have followed the Bosnia crisis quite closely. How is the situation in Kosovo different and how is it similar? The reader might have personal contacts in the area, or may have visited some of the geographical locations that the news story concerns, as illustrated by Dimitrova [1999]:

For example, receiving information that Topchider was bombed means something to me because I've been there. However, a CNN broadcast that states a part of Belgrade called Topchider was bombed means very little to a person who has no clue where Belgrade is and why the war is going on in there. The semantic filters working at the level of the consumer's mental frequency will have to account for personal background and preferences.

The division between these levels of news context is not clear-cut. News producer is often part of the larger community that it serves. This is typically the case with local newspapers. However, this is not true with CNN as it attempts to be a global news organization, and does not serve any communities specifically. Also news producer can be the community itself, as it produces information for all of its members. The discussion on the relationships between levels of context is further elaborated in the following chapter, and a prototype system for customized financial news is introduced in chapter 6.4.2.

2.3 Conclusions on the example services

The examples analyzed in chapter 2 - My Yahoo! and CNN Interactive's Kosovo theme - show that news stories can be efficiently filtered and presented to the individual customers automatically, and that stories can be augmented with illuminating background information using traditional journalistic work methods and by hyperlinks to connected material.

The main conclusions that can be drawn from these widely used Web news services motivate the rest of this thesis:

- customization is typically considered too narrowly to cover just information filtering for an individual customer; this work expands this view toward news augmentation and communities,
- customized products fall short in adaptivity to changing user needs, flexibility for managing multimedia content and their metadata, and in understanding and benefiting from special nature of the individual and community context of each customer, and
- automatic news augmentation can be used to customize news content and to complement work done by professional journalists.

It is not realistic to expect that all augmented news services will be produced automatically, but currently the balance is on laborious hand-coding of interdependencies between news articles and related material. New tools should help in building cross-connected, annotated, and metadata-rich news databases that can be more efficiently used for tailored news services.

The augmentations on CNN.com do not make a fundamental difference to the subjective experience and connection the reader has with the story. The idea in customized augmentation is to include additional material that is not only related, but also significantly better engages the news customer.

3 Understanding news content

This chapter presents some of the key characteristics of generation, delivery and consumption of news, and sets the ground for discussion on how news can be customized to prevent information overload and information decontextualization.

A *genre* is a stream of steady flow of materials that can play a definite role in the activities of some community. Each genre, like the news media, implies a relationship between the producer and the consumer of the materials in question [Brown and Duguid, 1996; Jones, 1998]. Genres permit people to seek out "more like that one", and to establish efficient, familiar, and habitual routines for using the materials that belong to a genre. *News genre* is one of the few original contributions of the mass media to the range of cultural forms of expression, and it is the core activity according to which a large part of the journalistic occupation defines itself [McQuail, 1993].

Our conception of news results from ancient way of storytelling [Darnton, 1975]. The main function of narratives is to make sense of reports of experience. They help to provide the logic of human motive, which makes sense of fragmentary observations, whether fictional or realistic [McQuail, 1993]. In one way or another, most media content is storytelling, which take rather patterned and predictable forms. Journalists are storytellers, and news accounts are typically cast in narrative form, with principal and minor actors, connected sequences, heroes and villains, beginning, middle and end, signaling of dramatic turns and reliance on familiar plots.

News results always from active selection and creation of stories by news reporters into the processed reality of news customers. This construction happens in three steps: (1) selecting what gets covered using the criteria of *newsworthiness*, (2) deciding what will become the focus of the story, i.e. determining the *'hook' or viewpoint*, and (3) deciding how the story is told. Thus, news does not really reflect objective reality, because it is a product constructed by journalists:

News is not something that happens; instead, news is what gets presented [Potter, 1998].

Different media provide news in a manner that best suits their technical capacities, evolved production methods, and consumption habits of customers. Print presents news in text and images inviting the reader to rebuild the scene of the events mentally, whereas radio and TV are more instant and continuous, and deliver fast on-the-spot reporting with a strong feeling of connection to news events [de Kerckhove, 1997].

The Web has developed into a major news medium in the 1990's. It has the following distinguishing characteristics as a news medium:

- multimedia content creation: multiple media journalism, requiring new talents from news producers,
- content customization: tailoring contents based on personal and community profiles,
- *flexible packaging*: versatile formatting and transfer of content to any medium, any delivery channel,

- anytime, anywhere delivery: ubiquitous access to content, especially through mobile devices, and
- evolving channel: new, unpredictable methods for news presentation are possible.

3.1 News content

News journalism includes multiple kinds of treatments of events: breaking stories, follow-ups, news analysis, interpretation and background, critical commentary, editorials, television and radio coverage, documentary, and book-length journalism. However, following and reporting the latest events (breaking news, the news flash, the news bulletin) are the main form of news journalism. Jack Driscoll [1997], the former editor-in-chief of the Boston Globe, broadly defines news as stories about events that are new, timely, and relevant to the readers. News involves public's right and need to know about events of general importance. News stories can inform, educate, guide and entertain the reader. Jack Fuller provides the following definition for news:

News is a report of what a news organization has recently learned about matters of some significance or interest to the specific community that news organization served [Fuller, 1996].

Journalists create, organize, filter and prioritize a continuous flow of information about on-going events, and then create a balanced *news package* for the readers or viewers from those information pieces. The *gatekeeper theory* has been used to describe the power of news editors, highlighting particular stories, promoting trends, deciding what information will be labeled meaningful, what will be sent further in a chain of information, or what will be saved for later dissemination [Shoemaker, 1991].

A large number of stories are rejected in news organizations, because they do not meet the needed criteria of importance and due to lack of space in the media. How are events valued? What is the mechanism for selection of news content that raises stories above the "news threshold"? News genre has generated over time a set of shared news values, or consensus, about newsworthiness of events. The following analysis (Table 1) uses a categorization of these news values by Hartley [1982].

Tal	ble	1.	Ν	ews	val	ues	[Harti	ley,	1982]
-----	-----	----	---	-----	-----	-----	--------	------	-------

Frequency	Relatively short time-span taken by an event.
Size threshold	The scale of importance of an event.
Unambiguity	The clarity of an event.
Meaningfulness	The cultural proximity or relevance of an event.
Consonance	The desire for predictability of an event.
Unexpectedness	The rarity and unpredictability of an event, balancing consonance.
Continuity	The 'running story' has continuing coverage.
Composition	The mixture of different kinds of events should be balanced.
Reference to persons	The 'newsworthy' people are followed.
Negativity	Bad news is good news.

There is a threshold below which an event will not be reported that varies depending on expected customer community, for example local and national news. Web publishing significantly lowers the size threshold. Also, "newsworthiness" can be defined as what is important and meaningful to the individual or community. The expression of a news message can be customized, and that complements the decision-making of human editors.

"Newness" of unexpected events usually gets discovered in thoroughly familiar and expected contexts [Hartley, 1982]. The Web allows collection of themes using news content that is inter-related as the base material. This can also help in bridging discontinuities in reporting.

Typically news is delivered in a package, where a particular type of *balance* is expected. This gives the customer an illusion of a thorough picture of world events. The task of attractive and balanced packaging, or a "news mix", is extremely difficult to achieve in an automated and customized system. This is where most of current systems for news customization fall short of expectations.

Events are often seen as the actions of people as individuals. Individual people are easier to identify than structures, forces or institutions. People-centric reporting relates to themes and continuity. It is relatively easy to gather large information packages about people appearing in the news. This is shown in the innumerable fan sites that have appeared on the Web. People are spontaneously gathering and publishing collections of information on other people they consider important. This is also a good example of a changing relationship between information producers and information consumers.

For the purposes of customization, news content can be analyzed along three axes: *depth, breadth*, and *distance* (Figure 11).



Figure 11. Three dimensions of news reporting: depth, breadth, distance

Depth

Depth means the thoroughness, understanding and commitment of the journalist to the story – shortly, the quality of the reporting. Newspaper usually gives more depth than television reporting. News journalists have different levels of skills and resources to describe and explain an event. New York Times generally provides better quality reporting than USA Today. Le Monde is more thorough and accurate than evening tabloid France-Soir.

Instead of telling the mere facts – *who, what, when, where* – of the news story, *why and how* search for the deeper underlying factors which lie behind the surfaces of the story. *Who, what, when, where* include the relatively transparent phenomena on the surface, *why and how* try to make things sensible, coherent, and explicable. People need not only know, but also to understand, the events in the news. To have an understanding of the events depicted in the news depends upon the depth in the news story, but this depth is rarely achievable [Carey, 1986]. Also, many problems exist in asking journalists to put more context into the stories: (1) placing story in context requires talent and experience to dig out great amounts of relevant contextual information before deadline, (2) journalists are given a lot of power when they can define the meaning of the event for the readers, and (3) there might not be enough space or airtime in the publication for contextual elements. Thorough coverage can be presented on the Web in addition to the ongoing news flow, since there is no limitation on daily print space or airtime. The only real limitation is the news customer's attention and ability to comprehend the news.

How can we increase the depth of news stories by customization? A story can be *explained* differently depending on what is known about the recipient of the message. Also, the level of expectations for background needed to understand these explanations can be set higher or lower, when news stories are customized. As an example, a story can point to previous articles that have touched on the subject. The mode of presentation of these related articles can depend on (1) how much the reader has followed the topic before, (2) how interesting the reader has found the topic, and (3) what is the expected level of knowledge of most important concepts related to the news story. Another example approach might be to look for similar stories or cases that have touched the communities the reader is a member of, and use these as a way to explain similarities and differences between new event and previously occurred events.

Breadth

Breadth means coverage of the news event from different perspectives. Background information broadens the scope of the event, and the event can be reported from multiple angles. This can mean opinions of different parties involved, or different domain experts. It can also be discussion amongst interested parties or interested readers on the topic. Multiple stories about the same event from different sources can be contrasted and presented to the reader.

The breadth of the news coverage can be added with customization by enabling two-way communication amongst the news customers on the topics, and by filtering most relevant comments for readers using information filtering technologies. The news customer might also have a preference of sources for different takes on the story from a local news source, national newspaper, newswire services, and global news organizations. Thus, the goal can be to increase the breadth by showing more connections and to understand the multifaceted nature of the event. On the other hand, breadth can be decreased through customization, for example by choosing only comments from those community members that the customer has valued highly in the past.

Distance

The distance of news reporting indicates the manner in which the topic is presented. A story can be told in a way that presents the events at a generalized and global level. On the other hand, the story can be *brought closer* to the listener, for example, by providing *example cases* of individuals affected by the topic of the news story. The story is *localized* using comparison with the living environment of expected audience. A simple example is how the journalists convert monetary sums to the local currency.

Personification and localization typically aim to help people to make the cognitive and emotional connection to the people and places behind the newsworthy events. The distance between the event and the reader can be shortened using mechanisms of customization. The story can be compared with communities more familiar to the reader, like the hometown [Elo, 1996]. Concrete examples of news content customization are presented in chapter 5.

3.2 Uses for news

Jensen [1986] has studied why people want news in the context of their everyday life. In general, news is expected to be a description of the events of a certain day, and people are oriented towards news with the motivation of making sense of the world. Four main uses for news can be identified [Jensen, 1986; Saari, 1998]:

- 1. *Contextual use*. Mass media news brings rhythm and routine to one's everyday life. Newspaper is read at the breakfast table, and evening TV news is watched when eating meal after workday. Although the news customers wants to focus on the news, they are deeply intertwined in a multitasking environment at home or at work. Typically, the news reception becomes semi-attentive and news is considered as peripheral source of information. News becomes background noise.
- 2. *Informational use*. People use news as a tool of being "updated" with the events in the world. People feel that they need the information mediated through news in one's role as an employee, a consumer, or a citizen in a society. However, people rarely check the validity of news and don't bother to check whether the "update" they have received is correct. Few people actually reflect critically on news and the informational uses of news can be quite superficial, and the updating is done for the purpose of re-

enforcing the status quo in one's life: nothing fundamentally catastrophic has happened so the person can continue with his life. News is a source of reassurance that the world is "still in place".

- 3. *Formation of social identity*. News can be a chance to feel commonality with other community members by raising awareness of current events, and offering common topics for discussion. The motivation to follow and understand news comes from assuring that an individual is able to participate and contribute to joint dialogue between community members. News builds and fosters communities and helps in socialization people know what to talk about.
- 4. *Entertainment*. News can be seen as just another way of killing time. Especially TV news are made to entertain the viewer, and there is a feeling of "watching reality" with a great potential to invoking emotional reactions in the viewer. Printed news rarely has the same effect. News is watched for drama and excitement.

Thus, there are many reasons why people follow the news, and only some of them are geared towards learning (i.e. deeply understanding what is going on in the world), and creation of new knowledge (i.e. seeing the meaning and reasons behind the news events and correlating them to personal experience) [Saari, 1998]. Neuman [1986] calls this casual semi-attentive media behavior *low-salience learning*. Typical news customer is not passive, but is somewhere in the middle on a scale from passive to active. The user's behavior is voluntary, and individuals actively scan, select, and interpret the flow of information that comes their way. The mindset of a news customer is typically semi-attentive and the news sources are considered as peripheral. Also the continuous accumulation of information and ideas from the media is typically a casual semi-attentive process. Reading newspapers and watching television are associated with relaxation and amusement. The distinction between seeking and needing information and browsing or being entertained is blurred [Neuman, 1991].

3.3 Community-centric news

Community-centric reporting most often manifests itself as local news, but communities are not only defined by physical proximity. A news producer typically *reflects* a specific audience. It shares with its customers a sensibility and a set of interests, tastes, values, and demographics. News journalists get feedback from the communities they serve in many ways; letters to the editor, phone calls from the readers, comments from colleagues, popularity and sales statistics, online discussions, and direct email.

News producers not only serve but also create and maintain their communities. The big challenge for news journalism is to simultaneously support the identified specific characteristics of different communities, and the heterogeneity and changing nature of their interests and needs.

On the Web, anyone can publish information independently at any time. The current proliferation of personal homepages and community sites shows that people do not hesitate to publish their own content. Individuals and communities can easily construct stories and messages, and to communicate with their

physical and virtual communities. Communities can become active providers of content, according to the model of *many-to-many communications* of newsgroups, mailing lists, chat, special interest web sites, and other Internet-based communication systems. This means a rapid increase of volume and diversity of online information and constant changes of information sources in number, content, and location. As a result, the validity of old journalistic practices, especially newspaper journalism, has been questioned [Katz, 1994]. As the journalistic "gatekeeping" becomes an option instead of a necessity, the journalistic profession needs to re-position itself by proving its added value by having a more active role towards its customers.

The Internet equalizes the possibilities in terms of access and distribution, as media content production used to be exclusively in the hands of the professional communicators. Communities formed around specific interests have their own newspapers and other news sources, like web sites. However, community-oriented news content online has been closer to the chatter or gossip group than to the formal quality of objectivity associated with traditional news sources. However, many "grassroots activity" Web sites already show how community publishing can result in a meaningful and entertaining addition to traditional publishing.

Silver Stringers ⁴ and their Web publication, *The Melrose Mirror*, is an example of a community of practice that actively writes stories about their topics of interest [Driscoll et al., 1997]. This electronic publishing community consists of about 35 members aged between 68 and 88, who regularly meet at the Milano Senior Center in Melrose, MA, to discuss about the contents and design of their Web site. This collaborative enterprise consists of several hundreds of items (stories, images, and music) created and maintained by the community members.

Most media companies and government regulators still continue to treat the general public only as consumers, rather than as potential producers and designers of content. However, the balance between information suppliers and consumers in networked media is shifting.

3.4 News contexts

The classification into event, producer, community, and individual contexts introduced in chapter 2.2, serves as basis for the discussion of decontextualization and recontextualization of news content (Figure 12).

⁴ Available at http://stringers.media.mit.edu/



Figure 12. Different levels of context in news creation and reception

Event context

A news story creates a triangular relationship between the event, the communicator, and the receiver (Figure 13). The communicator of the message (the journalist) must not only have an orientation towards the event itself, but also an orientation towards the receiver of the message [Hartley, 1982].

Personal storytelling relationship between the teller and the listener has been unachievable in broadcast communication. As Hartley [1982] notes:

When we turn from personal relationship to the relationships between newspapers or TV and their audiences, we are faced with a problem. They too must include an orientation to their addressees as part of the message, but except in the most general and abstract of ways they don't know who the addressee is. It is the problem media cannot escape. They **must** develop a practical 'mode of address' which expresses not only the content of the events they relate but also their orientation towards the viewer or reader, since this orientation is an unavoidable constituent element of all language.

In digital media, a personal relationship between the storyteller and the addressee can be created. The problem is that there are not enough journalists to tell the story personally to each potential listener, so the model of personalized storytelling in news reporting depends on customizing a generic story (or a limited number of variations of a story) computationally at the time of reception.



Figure 13. Triangular relationship between the event, communicator, and receiver [Hartley, 1982]

Story points to an event at a specific space and time. This spatio-temporal dependency of news stories is a fundamental feature of news content. News journalism emphasizes the recent event at the expense of other events which occurred before or are already known. Journalists recognize the need to include *background information* about the event into the stories, but it is rare that a news story contains more of it than is necessary to permit the assumed reader to make some sense of the new material. News journalists do not typically try to teach people about the topics in the news.

The role of customization is to analyze the event and its context and to decide if the news story is relevant to a reader and how the event could be made more understandable, for example by comparing the event context with the community context of the customer.

Producer context

Judgments of news value are often relative and based on journalistic "feel for the news" at the particular moment based on the "news sense" that the journalist has developed over time. Media organizations are both mirroring and making the social criteria that form the basis for newsworthiness. Traditional media products are very well defined in both their style of content, style of presentation, and the amount that can be presented or printed, as only a certain number of events that are considered newsworthy.
The news stories have internal and external origins. A typical news reporting process goes as follows. The ideas for stories are discussed in a news conference at the production location. The stories are assigned to reporters. Reporters interview sources, follow leads, and check facts. News wire services are followed for new story ideas and breaking events. The stories are chosen or discarded at the *news desk*. Layout/lineup of the stories is decided before final editing to produce the content page for a newspaper or an anchor script for the broadcast news. Therefore, in constructing the stories journalists are often strongly influenced by elements outside the event, such as deadlines and space limitations.

The rules of a news genre are not codified, as past instances of news stories create precedents and expectations for the interpretation of subsequent stories. This creates a pressure for future communications to conform to the pattern established by earlier ones. The selections and construction decisions for news are made from *news perspective* of the news production organization. News perspective explains how the news staff selects and treats the news to get their work done. It is a sort of *professional news filter* or bias that help journalists simplify and organize the overwhelming amount of material they must sift through [Altheide, 1976]. Mass media's evolved preference for certain types of events, issues, and people, is often criticized as *agenda setting*.

Timeliness is an essential ingredient of both novelty and relevance, both of which are highly prized in news. Pierre Bourdieu [1998] argues that news organizations often write their stories as a reaction to their competitors' content. There is a constant race between competing news producers to find the latest information on a given topic that sets them apart from their competitors. Bourdieu writes that "*this game of mirrors produces a formidable effect of mental closure*" as news sources become more and more similar in content. The importance of being first - *to scoop* - is often greater than that of any other factor in deciding on selection and prominence.

It is an interesting question whether this producer context can be somehow filtered, or made more apparent, through customization. If the news story is accompanied with links to other related articles, the producer context "bias" can be maybe more easily detected. This process can be aided with automatic systems for bias detection [Sack, 1994].

Another way of looking at producer context is that media organizations possess large digital content warehouses that can be flexibly packaged and searched online. Reporters can routinely check "what we have written about this in the past". If these databases are well structured and maintained, they become an important asset for the media organization. They also provide natural content for augmenting on-going events with related information.

Community context

News producers target their stories toward *an audience* made of individuals who can belong simultaneously to many communities. These communities are often fertile ground for discussing news events, negotiating shared meaning for information, and for creating new knowledge for the benefit of the

community based on news stories. Communities use the Web for discussion, but also for filtering of content, and for own community publishing activities.

For purposes of this work, three kinds of communities are identified:

- *community of proximity:* people living or working in a certain area, a family, an organization (local news, organizational newsletter),
- *community of interest:* people with common interests around a topic area (hobby, fan magazine, professional journal, mailing list), and
- *community of practice:* people with joint enterprise and actions, as engaged members of the community are known by other members.

There is no clear division between these types of communities, as a community of interest is often also a community of practice, and local communities tend to become communities of practice. Often the producer also belongs to the same communities as the readers.

The producer context and the event context are often part of the community context that they serve. For example, a local newspaper has been traditionally the cornerstone of community. It is considered as an essential part of building community identity. Similarly, the local online newspaper can be a gateway to the community, possibly also becoming the default Web news page for people living there. It does not just provide the news, but through additional services (such as directories) helps local readers to discover their community. City guides on the Web are a good example of content and service aggregation. These city guides offer a large combination of local information and services, chat rooms, classifieds, yellow pages, maps, travel agents, ticketing, and many more.

Local newspapers rely heavily on wire services such as AP or Reuters for news outside their community. However, these wire services do not indicate the relevancy of their articles to the readership's community [Elo, 1995]. Community news source thus 'localizes' the article, using knowledge and insight on how to illuminate the story from a local point of view.

Communities of interest and communities of practice have much less clearly defined boundaries than communities of proximity. They also have often substantially smaller number of members, than the geographical communities served by local news sources. As a result, there has not existed a good economic or technical model for serving these communities. By using mechanisms of community modeling, which capture the identity and interests of each community, content can be targeted to fairly small groups efficiently. This is where customized news services have large potential.

Individual context

Individual news customers have different interests, expertise, background, and goals in their usage of news in different contexts and multiple roles that people have in their daily lives. Our experience of the world results in an incrementally created and continuously evolving structures of knowledge in each individual.

The dissimilarity of these structures is one of the key reasons for individual differences in how we perceive the world. This information base is what we use for context when interpreting news. For example, events that are common in individual context are understood better than the events outside of personal experience.

Personalized news services can assist an individual in noticing and better understanding news events that are most relevant to that person. Personalization always requires some kind of user model, which is used as a basis for tailoring the message. The user model can consist of information of possibly multiple roles of an individual, including personal interests and expertise. The possibilities for modeling individual news customers and the context of their media usage are further discussed in chapter 6.1.

3.5 Decontextualized news

The problem of decontextualization of news relates strongly to the historical development of news content, and has gradually progressed with the invention of new technologies [Postman, 1985]. The telegraph allowed newspaper to operate in "real time" and turned competition among newspapers onto timeliness. The language of news was flattened and standardized. *The pyramid style* of reporting was formed: most important facts were transmitted first because of the unreliability of the communication channels [Fuller, 1996]. A "stringer" replaced the traditional news correspondent, and supplied the bare facts into the newsroom for another person at the other end of the telegraphic line, who became responsible for reconstituting the story.

Because the number of events that had to be analyzed at the newsroom grew, a more efficient routine of news handling had to be put in place:

News became treated as a commodity: something that could be transported, measured, reduced, and timed. This change in the production routines divorced news from an ideological context that would explain and give significance to events. By elevating objectivity and facticity into cardinal principles, the press abandoned explanation as a primary goal. The announcement of an event was divorced from its analysis and required the reader to maintain constant vigilance to the news if he was to understand anything [Carey, 1986].

Another reason for decontextualization of information is the commercialization of media and the search for largest audience possible. Neil Postman [1985] claims that we follow all media, also news, mainly for entertainment. Television is a powerful medium in providing drama and emotion, but the visually rich and linear style of storytelling is often not combined with a capacity to inform in depth about complicated matters. Postman describes the world of television news as "*the peek-a-boo world*" – event comes and goes, pops into view for a moment, then vanishes again. He argues that television alters the meaning of "being informed". People follow *infotainment* that creates the illusion of knowing something but which in fact may lead one away from knowing.

In a study of recall in U.S. television news, subjects were asked what they could recall of news broadcast they had just seen [Neuman, 1991]:

On average, respondents could recall one news story out of 20 without additional prompting. When they were asked for a list of headlines from the news programs they had watched, they could recall four additional stories and provide details about those stories. They claimed to recall another four stories, but could not remember any of the details. So as a rough rule of thumb, 5% of the stories are recalled unaided, and another 45% with prompting, only half of which are substantiated with supporting details. However it is not clear if these statistics change, when people watch multiple newscasts the same day, and the message is reinforced.

This problem can be traced down to three factors [Levy, 1992; Gunter, 1987; Bourdieu, 1998]: (1) most news stories on television or radio do not contain much information, (2) people who rely primarily on television news often do not get news from other sources, and (3) most stories are presented with little context that would help viewers understand the story.

Typical news stories are limited to basic facts of an event, and do not help news customers to construct the meaning of that event. This applies to both television and newspapers. Decontextualization means that *news faces a problem of comprehension*, and that there is a need for instructive mediation between the news provider and the news recipient. One possible key to coping with growing amounts of information is to understand the topics of interest in the news at a more complete level.

With the Web, the amount of available news content is growing extremely rapidly. As any content becomes accessible through a hyperlink or a search engine, the volume of news-related data can become overwhelming. Online publishers are seeking for more depth of content in integrating archives of historical articles and other background material with the continuous news flow. However, in the world of abundance of easily accessible information, the need for orientation and guidance becomes more apparent [Wiley, 1998]. Fuller writes [1996]:

Information overload manifests itself emotionally as a profound sense of loss of control. Today the issues that shape our future have become so complex that they sometimes seem simply too hard to understand even at the most basic level, let alone locate them in a context that lets us exercise control over them. Loss of context seems to be more important problem than the share amount of information.

3.6 Summary of findings

News has typically been offered as a ready-made package that the journalists produce according to the production rhythm of their media. Digital media and the Web are changing this in three important ways: (1) news content can be tailored dynamically to individuals and communities, (2) communities are producing the content themselves, which is equally accessible as the content of media companies, thus challenging the professional journalists to find new ways to make relationships between media producers and active communities, and (3) without carefully designed news services, the customer will likely suffer from the effects of information overload and information decontextualization.

Customization can help to increase the depth of the story by making the events behind news more understandable, to tailor the breadth of reporting to the needs of the customer, and to reduce the perceived distance between the news customer and the event depicted in the news. Of the examples mentioned in chapter 2, My Yahoo! is decreasing the breadth of overall scope of reported events through information filtering, and simultaneously increasing breadth of issues that are known to be relevant for the customer. The Kosovo theme site from CNN.com increases both depth and breadth of the story, but does not do that in a customized way. Neither of these services shortens the distance between the news event and the reader.

Research has shown that current news media often fail to make events understandable to their customers, as news producers emphasize speed and entertainment value in reporting. How could this computational teacher help user's understanding of complex concepts inherent in everyday news topics? Can an automated news augmentation system take the educator's role that journalists hesitate to take? Could the machine understand the extent of user's domain knowledge? Before going deeper into these topics, I discuss some important factors of learning and creation of new knowledge as they relate to news.

4 Learning and knowledge creation through news

The purpose of this chapter is to discuss the relevance of cognitive theories of story-based and situated knowledge construction and knowledge creation in relation to design of systems that produce customized content. As mentioned in the previous chapter, current news media face a *problem of comprehension*. News is not presented in a manner that is optimal for learning and understanding. The role of the news media has shifted away from educating people about the world to intense competition amongst the news producers on *who is first* and *who gets people's attention*. This has resulted in news services that are comprised of largely decontextualized information flows that people follow, not for the sake of understanding, but for "fast updating" and entertainment.

People need to develop special *media literacy skills* to really be able to understand what the events depicted in the news are about. Potter [1998] has summarized different skills and required knowledge to increase the ability to understand the news (Table 2). One purpose of this work is to make new tools that help people develop these skills easier and faster than before. How can we make news systems that invite to explore and understand, not just to watch and follow?

	SKILL	KNOWLEDGE
COGNITIVE	Analyzing the important points of a storyDetermine if the story is balanced	 Historical context from previous news stories Knowledge about the topic from real-world sources and personal experiences
EMOTIONAL	 Ability to put one's self into the position of different people in the story Extend empathy to other people in a similar situation to the news story 	 Personal experience of similar situations Personal links to event context (people/places in the news)
AESTHETIC	Analysis of the visual syntaxAnalysis of the narrative structure	 Knowing the visual and audio elements available for news producers Knowing what narratives are used in news stories
MORAL	 Identifying your personal moral stance in reacting to the people and events in the story. Separating people from actions Determining relative responsibilities of actions Ability to consider extenuating circumstances 	 Accessing information about histories of the featured people Having information about similar events to use as a template for this story

Table 2. Media skills and knowledge related to news [Potter, 1998]

What are the reasons behind our inability to understand and remember news? First reason is the fast pace of news reporting and emphasis on the present events. The rapid speed and brevity of news presentation hinders absorption of stories into the long-term memory, thus preventing the creation of new knowledge. The news story cannot be connected with already absorbed information, although the related knowledge from past news events might be remembered by the individual, and could be activated with better cues for remembering [Graber, 1989]. The other problem with learning from news is that the messages that comprise the news are too far removed from the individual and community context of news customers, as discussed in the previous chapter.

4.1 Schemata, stories, and scripts

In his book *Remembering*, Bartlett [1932] reported a series of memory studies that used a broad range of material, including text and folktales. He found that individuals recalling narrative material made inferences and other changes that lead to more coherent and concise story. The nature of these changes was more conventional than what could be found in the original stories. He concluded that human memory is not a reproductive, but a reconstructive process. Bartlett introduced the concept of *schema*, meaning unconscious mental structures of human knowledge that capture the generic aspects of the world [Brewer and Nakamura, 1984]. He argued that the changes found in story recall could be accounted for by assuming that schemata operate on new incoming information to fill in gaps and rationalize the resulting memory representation. Although almost completely ignored in Bartlett's lifetime, the schema construct has been re-adapted widely in psychological research on human memory and in the field of AI.

Marvin Minsky introduced the notion of *frame-based systems* [Minsky, 1975], which were significantly influenced by Bartlett's theory of schemas. As a knowledge representation mechanism, Minsky's frames are knowledge structures that contain fixed structural information. A frame represents a concept or a situation. These frames have slots that accept a range of values; each slot has a default value that is used if no value has been provided from the external world. Attached to a frame are several kinds of information, for instance, definitional and descriptive information.

Schank and Abelson [1977] proposed a theory of cognition that is based on human ability to understand and tell stories. A *story* is our means to communicate our thoughts and intentions to others. A *script* is a set of expectations about what will happen next in a well understood situation:

Understanding means mapping your stories onto my stories. Thinking means finding the right script to use, rather than generating new ideas and questions, so, essentially, we find it easier to apply scripts than to reason out every new situation from scratch. People need a context to help them relate what they have heard to what they already know. We understand events in terms of other events we have already understood [Schank and Abelson, 1995].

What is this reminding process? Each story we hear reminds us of one that we know, and in a conversational situation we tell that story if it seems appropriate to the audience, adapting it as we go to make it more relevant. Finding a relevant past experience that will help make sense of a new experience is at the core of intelligent behavior, and reminding is the mind's method of coordinating past events with current events to enable generalization and prediction [Schank and Abelson, 1995].

Scripts can be considered as a sub-class of schemata that are used to account for generic (or stereotyped) sequences of actions. The sets of facts are organized by categories and the category is defined as a set of features that serves as base for inferring membership. Graber [1989] defines *schemata as cognitive structures consisting of organized knowledge about situations and individuals that has been abstracted from prior experiences*. These schemata, similarly to Schank's scripts and indexes, provide guides to selection, relevance and cognition. Schemata are used in processing new information and retrieving stored information. They also help in evaluating information and filling gaps when information is missing or ambiguous. The more knowledge we have in a particular domain, the easier it is to process new information about that domain and the more likely it will be remembered [Potter, 1998]. Schemata have been applied to explain the interpretation and influence of news [Graber, 1989; Höijer, 1992; Tapper, 1995; Saari, 1998].

People thus evaluate news based on prior experience, with clear emphasis on personal experiences that relate to the events in the news [Graber, 1989]. Therefore thinking involves some form of *story indexing*. In order to assimilate a story, we attach it into our memory. In this process, concept formation and categorization plays an important role. Also, the more contextual information we are given about a situation, the more places we can attach it in our memory, and the more ways it can be compared to other events.

How does this relate to news customization? Stories, scripts and schemata are important in both media consumption and media production. If the news can remind us of something we have already understood, the new message will also be better understood. In addition, the journalists have their background experience and their individual story index, which influences the way news is reported, how the new story will be told, and how the recipient will interpret the story [Bourdieu, 1998].

Stories are often *elaborated* for a stronger reaction in the hearer using detail addition and commentary [Schank and Abelson, 1995]:

• *Detail addition* makes a story more vivid or realistic. Although the addition of details may not significantly change the story content, they are important in terms of storytelling. The more details the teller adds to the story, the more memorable the story becomes. Thus, details make the story interesting and attention-getting.

• *Commentary* means adding the teller's own view of the situation, including comments on how well or how poorly various people behaved, what the right thing to do would have been, or the others might say or think about the situation, what we would do the next time, and so on.

News augmentation can include these two forms of story elaboration: details about previous events and comparative statistics are added to the news story, and the news story is combined with commentary from other news sources, other readers, and community members.

In personal communications, we adapt the story to the listener. When a person tells about a news story to a friend, the structure and the point of view of the story are chosen based on what the teller already knows about friend's interests. The "optimal" story would always lead the friend smoothly from point to point, informing of the salient facts and important arguments. This is a fundamental part of social interaction. However, traditional news media cannot exploit this natural adaptation inherent in storytelling. Research has shown that people sub-consciously treat media – as well as computers – like they treat humans in a social situation [Reeves and Nass, 1996]. To conclude, if computers could adapt the story to the receiver, that would be acceptable social behavior, and content customization is a step toward more adaptive storytelling media.

4.2 Conceptual models for sensemaking

Several areas of research – digital libraries, information retrieval, information integration, and Internet information access researchers – have worked on methods and tools, which aim at description of knowledge in a formal model. However, it is clear that there will never be a conceptual model or categorization that would satisfy the requirements, needs, and different viewpoints of all people. Conceptual models are always bound to be fluid and negotiated. Wittgenstein [1953] argues that categories exist insofar as people *use* them, and that they are bounded to the extent to which individuals in a community can agree on category boundaries. However, although no universal classification scheme exists, the role of categorization is central to understanding, and plays a significant role in the information environment of the implemented news filtering and news augmentation systems described later.

In the context of this work, structured descriptive information about information is needed because content-related knowledge is best represented in the form of conceptual categories. This categorization attached to information artifacts helps information explorers to move between the document space and the concept space. The *document space* means information artifacts that have an external identity in the world. Information artifacts are distinguished from the conceptual knowledge that arises when a person interacts with these artifacts. The *concept space* is the cognitive realm of the context and knowledge required to interpret documents. This definition includes the idea of concept space as a hierarchical organization of topics [Baldonado, 1997]. Explicit and structured classification systems of shared conceptual knowledge, like the Dewey Decimal for libraries, or Yahoo! category for Web resource classification, attempt to build

a bridge between document space and concept space by communicating a conceptual framework of reference [Baldonado, 1997].

4.2.1 Ontologies

Both *ontologies* for automatically processable, knowledge-rich content characterization of information sources, and *metadata* for information source description, evaluation, and access, have an important role in scenarios for intelligent information access. In almost all news services, the content is organized according to some form of categorization. The value of good-quality and detailed categorization increases as the news is to be used in an asynchronous and customized manner.

Formally, in the context of knowledge representation, ontology is a partial specification of a conceptual vocabulary to be used for formulating knowledge-level theories about a domain of discourse; i.e., ontologies provide a way to formalize a common vocabulary for describing some area of interest [Gruber, 1993]. When agents (human or artificial) agree on such a vocabulary, then they can share and reuse knowledge.

In the context of this work, *domain ontology* means a set of formally specified conceptual structures. Domain ontology is used for modeling the semantics of the information content. Ontology can cover multiple *dimensions* of the content. By organizing a collection of documents along dimensions, the cataloguer provides users more choices to identify key concepts of documents. Typically a cataloguer works with a set of conceptual dimensions that have been defined and standardized, such as MARC (for library resources), and Dublin Core (for network resources) [Weibel et al., 1995]. These dimensions include not only subject, but also other dimensions such as author, publisher, and publication date. With these dimensions the ontology should be able to cover the semantic needs to produce and deliver customized content to the customer.

Ontology comprises a set of concepts and concept relationships representative to the problem domain (Figure 14). Concepts and their relations define *conceptual models* for classifying information objects under different *dimensions*. A dimension is used in this context to describe different aspects of the content domain. Another closely related term to metadata dimension, which is often used in connection to information retrieval, is *a facet*. Taylor [1992] defines facets as "clearly defined, mutually exclusive, and collectively exhaustive aspects, properties, or characteristics of a class or specific subject."

Metadata dimensions can be independent from each other, like *subject matter* and *media type* [Savia et al., 1998]. However there are often subtle interrelationships between these dimensions that can be difficult to model. For example, *geography* dimension consists of continents, countries, cities, towns, neighborhoods, and *industries* dimension might include a sub-model of *car industry* of car makes, manufacturers, and models. These two dimensions are not fully independent, because a car is manufactured in some geographic area, and this information may be useful in categorization (European car vs. Japanese car).



Figure 14. Different dimensions of domain ontology [see Part II: Publication 5]

When conceptual models are used in news customization, for example adaptive news filtering, interrelated dimensions significantly complicate the problem of automatically learning and adjusting user's interests and expertise. If a person buys a Japanese car, it rarely raises the buyer's interest in general news events happening in Japan. On the other hand, if the same person is interested in Japanese *companies* (another metadata dimension), the Japanese news events can be more interesting. In the SmartPush project a dimension is defined in such a matter that their concepts are independent of the concepts in the other dimensions [Savia et al., 1998].

A *concept* may be defined as a generalized idea, a mental entity, or a category. It can be used as a synonym for thought, reference, or meaning. *Categorization* refers to the process of dividing the world of experience into groups – or categories – whose members bear some perceived relation of similarity to each other. Categorization is fundamental because it permits us to understand and value predictions about objects and events in our world. Conceptual models can be used for knowledge organization, classification and information retrieval. They consist of concepts and relations between concepts, that is, semantic nets. Conceptual model consist of mappings between terms (word occurrences) to concepts. *Thesaurus* is a closely related term used for a documentation language and a repository of lexical semantics for Information Retrieval. *Entity-Relation (E-R) model* [Chen, 1976] and more recently *UML (Unified Modeling Language)* [Booch et al., 1997] are other often-used tools for defining complex semantic interrelationships among data elements, primarily in defining and modeling information systems.

4.2.2 Sensemaking

People must continually make sense of large amounts of information. *Sensemaking* is the process of searching for a representation and encoding data in that representation to answer task-specific questions [Russell et al., 1993]. Information explorers use interfaces and tools for finding documents in order to carry out a task in the world for which they do not have enough prior understanding or ready-to-hand concepts. Could a news service be designed to help the customer to better make sense of the news?

The ASK (*anomalous state of knowledge*) model of information exploration [Belkin *et al.*, 1982ab] assumes that people seek information in order to fill in holes in their knowledge structures. However, in any communicative act, both the initiator and the receiver construct and interpret the act according to their own prior contexts and backgrounds. It is through continued dialog and mutual commitments that the two parties arrive at a shared understanding [Baldonado, 1997].

A recurring pattern in sensemaking is the *learning loop complex*, which involves four steps [Russell et al., 1993]:

- *search for representations:* the user looks for good conceptual structures into which the information of interest for the task can be placed,
- instantiation of representations: the user fits the information into the chosen conceptual structure,
- *shift of representations:* if some important pieces of information do not fit into the conceptual structure, the representational structure is revised,
- *use of instantiated representations for task-specific purposes:* representation is used either as the input to the next step in the user's task or as the solution to the user's underlying problem.

Russell *et al.* [1997] emphasize the cyclic and iterative nature of this cognitive model. The evolution of a user's interests can be explained in this model as the movement of new concepts from the periphery of the user's attention to the center. The new understanding at the conceptual level has a direct influence on what the user looks for next in document space [Baldonado, 1997]. *SenseMaker* [Baldonado and Winograd, 1997] allows users to interactively build document-based structures in order to allow key concepts to emerge, and to build on and sculpt these interactively constructed structures in order to perform concept-based finding and filtering of documents.

From a cognitive perspective, tightly integrated links from collections of documents to concepts make it easier to identify and recognize key concepts. If the relevant conceptual structure is made explicit through the interaction, a user can better see how the document concept applies to these ontological structures [Baldonado, 1997]. This principle has been an important guideline in the implemented news augmentation system described in chapter 6.4.2.

To take the situated view of conceptualization, every categorization is with respect to the person's history. In people, conceptual categorization of role and identity within a community are overarching; hence experience is subjective [Clancey, 1997]. Conceptualization doesn't strictly map onto words, definitions,

and rules because many conceptualizations are nonverbal and because the meaning of even verbal conceptualizations is coupled to nonverbal categorizing.

Learning is an incremental process of building upon existing domain knowledge with new concepts about that domain. When learning about unfamiliar items in a specific domain, referencing other more familiar and related concepts makes learning the new concept easier. When teaching a new concept, the existing knowledge about domain concepts should be used to make comparisons to new concepts that are being introduced. Comparisons are often used in descriptions to reduce the cognitive load on the reader. A model of user's expertise in a domain can be employed to produce descriptions, which introduce new concepts by comparison, thus grounding descriptions to user's existing knowledge [Milosavljevic, 1997].

Fortunately, from the perspective of customized news services, news content consists mostly about concepts that fall into dimensions that can be defined in a conceptual model. Also to a certain extent, the role of the news industry is to create and maintain these models for the "domain of news", meaning the topics that get covered as news. In this sense, news media creates mental models or a "map of news landscape" for their customers. One role of customized media is to understand how this map can be changed to best meet each customer's needs.

4.3 Socially situated knowledge and expertise

Social networks can be created and supported by computer networks, as they connect people and organizations. In a networked society the boundaries of networks are permeable, interactions and linkages are encouraged, and hierarchies are flat. Another important property of communications in social networks is that interactions shift from those based on characteristics people are born with (age, gender, race and ethnicity) to characteristics they have adopted throughout their lives (lifestyle, shared norms and voluntary interests) [Wellman, 1999].

Nonaka and Takeuchi [1995] also emphasize the social aspect, as they define knowledge:

- Knowledge is about *beliefs* and *commitments*. It is a function of particular stance, perspective, or intention.
- Knowledge is *related to human action*. Information is a flow of messages, while knowledge is created by that very flow of information, anchored in the beliefs and commitment of its holder. It is always knowledge "to some end".
- Knowledge, like information, is about constructing *meaning*. Both information and knowledge are context-specific and relational, because they depend on the situation and are created dynamically in social interaction among people.

Expertise is socially contextualized if it reflects the community's activities of inventing, valuing, and interpreting theories, designs, and policies. This *expert-in-context* viewpoint emphasizes that the meaning of expert knowledge does not reside only in the individuals, but rather arises in a dynamic interaction

matrix involving the individual and his/her community context. Human minds develop and extend their sphere of activity and communicative competencies in social situations.

Situated storytelling is an important factor in learning and resolving problems facing communities of practice. In studying Xerox service repair technicians, Orr [1990] found out that technicians rarely relied on the company-provided service manuals when troubleshooting machine problems. Instead they used "war stories" passed from technician to technician in an oral storytelling culture. Orr pointed out the value of these stories in knowledge creation and sharing, noting that they represented an important intellectual resource that the company should capitalize upon.

As a result of their knowledge and previous experience of the world, individuals acquire in-built readiness to notice certain features of their situation significant. Attribution of meaning to these selective perceptions enables judgments to be made and intentions formed, which may lead to purposeful action to try to realize the intentions. Bateson [1973] names two kinds of learning (Learning I and Learning II). First kind is obtaining know-how in order to solve specific problems based on existing premises. The second kind of learning is establishing new premises (i.e., paradigms, schemata, mental models, or perspectives) to override the existing ones. The challenge for designers of communication media, such as online news, is to create environments that make Learning I and Learning II empowering and productive, and to design computer tools suited to human use in the context of *purposeful human activities* [Suchman, 1987].

4.3.1 Communities of practice

Most communities engage in some degree of collective cognition – the interactions through which they learn from one another's experiences, set common strategies, develop a shared vocabulary, have common norms and means, and evolve a distinctive and shared way of thinking [Agre, 1998]. In any *joint enterprise* - working, living, playing - people develop shared ways of doing things and relating to one another that allow them to achieve their joint purpose. Over time, the resulting common practice becomes a bond among the involved individuals. This social interaction is a critical component of *situated learning* as learners become involved in a *community of practice* [Lave and Wenger, 1991]. Situated learning is often unintentional rather than deliberate, and the communities that matter are not always the most easily identifiable, because they often remain informal. The individuals in these communities inform each other about interesting events, for example news that can have an influence on their shared practice.

Typically, people belong to many communities of practice at any given time. For example, a business organization is not a single unified entity, as it involves more than one practice, and thus is typically a constellation of interrelated communities of practice. The relation of individuals to organizations is therefore not direct, but takes place through participation in communities of practice.

The terms "community" and "learning" are often considered very positively, almost idealistically, as a situation where people coexist peacefully and mutually support each other, and learn to make the world a

better place. However, community is not good or bad in itself, as there are typically disagreements, tensions, and conflicts among the members of a community of practice, and the joint enterprise of the community of practice can also be harmful to the larger community [Wenger, 1998a]. We also learn all the time, whether or not we learn what is expected of us, or what is good for our communities or us.

Community of practice supposes that the members (1) are aware of the existence of the community, and (2) have created personal relationships with other members of the community. The community is not static - new members can arrive at any time and existing members can leave, but there is a sense of belonging to a group of individuals, and the group is relatively stable.

There are different types of membership to a community of practice: central or peripheral; old-timer or newcomer; central to member's identity or incidental; some community members are community innovators and are typically positioned as thought leaders. Whatever the form of membership, situated learning is both the vehicle and the result of participation in communities of practice, and it affects the identities of those involved by changing their sense of how they can engage with the world [Wenger, 1998a]. Membership is often a matter of competence, of being able to participate in socially defined activities and to contribute and engage in a community and its joint enterprise. In these communities of practice, knowing, belonging, and doing are not separable, but together form the experience of participation.

The reason for emphasizing communities of practice in this work is that news media have traditionally neglected them, because they have been difficult to identify and serve through mass media channels. However, they can be considered as important units of shared needs and joint interests that require special services from information providers. They are also active and creative entities that the media companies could benefit from.

4.3.2 Boundaries and membership

As communities of practice form, they create boundaries between those who have been engaged in the practice and those who have not. These boundaries are created by differences in the perspectives, the languages, and the styles that characterize each practice. Boundaries often confront newcomers or outsiders who seek entry into a community of practice, for example in being unable to participate in a conversation. As a newcomer moves from the periphery of a community to its center, they become more active and engaged within the culture and hence assume the role of expert or old-timer. Membership in these communities of practice is decided by individuals' abilities to trade practically valuable information [Brown and Duguid, 1991].

Community boundaries are reflected in our identities. Because we belong to many communities of practice, our learning includes the process of reconciling different forms of membership. Furthermore, boundaries have to be crossed for communities of practice to work together and for their various

perspectives to be coordinated. People who cross these boundaries have different experiences in different practices. They may or may not broker learning from one community into another. Thus, the boundaries of practice may either be liabilities or learning assets and fertile ground for innovation.

Successful communities of practice provide forms of participation that encompass the past and open the future. The quality of participation is not just defined in terms of activities, but as a trajectory through time [Wenger, 1998a]. Sustaining a practice over time provides both depth of knowledge and the potential for the creation of new knowledge. Figure 15 mirrors the typical stages of development of a community of practice.



Figure 15. Typical lifecycle of a community of practice [Wenger, 1998b]

Wenger [1998a] emphasizes that membership in a community of practice is a matter of *mutual engagement*. A community of practice is not just an aggregate of people defined by some characteristics. Thus, homogeneity is not a requirement, nor the result of, the development of a community of practice.

What makes engagement in practice possible and productive is as much a matter of diversity as it is a matter of homogeneity [Wenger, 1998a].

This emphasizes the role of communities of practice with a unique profile and as a combination of different skills of the individuals. This is an important point when considering the possibilities for news customization. A community is not just aggregate or average of some characteristics of its members. Therefore, there is a need for *community modeling* in addition to modeling the individual members of communities.

Wenger [1998a] denotes that negotiation of meaning is a fundamental process in social learning:

All that we do and say may refer to what has been done and said in the past, and yet we produce again a new situation, an impression, an experience: we produce meanings that extend, redirect, dismiss, reinterpret, modify, confirm – in a word, negotiate anew – the histories of meanings of which they are part.

Knowledge management has a strong emphasis on information capture and reuse. However, information to be used as memory is valuable only when it can be contextualized, decontextualized, and recontextualized at the proper time. As information propagates across individuals, intra-organizational, and interorganizational boundaries, it necessarily loses some of its context. *Boundary objects* in an organization work because they necessarily contain sufficient detail to be understandable, but at the same time, neither party must understand the full context of use by the other. As boundary objects, artifacts and other potential resources are given to other people. This requires the information to be decontextualized. Otherwise, the secondary users will drown in unnecessary, unhelpful, or conflicting data. Without shared meanings or histories, information will lose context as it crosses boundaries between communities of practice. Sometimes this loss is beneficial, in that it hides the unnecessary details of others' work. More often, however, information must be recontextualized in order to reuse experience or knowledge [Ackerman, 1998].

There are coexisting requirement for contextualization, decontextualization and recontextualization. To use information as a memory, one must remove the detail that provides context, making the information into a boundary object. However, at the same time one must consider how others will use it later as a resource in their processes; otherwise, subsequent users of the memory will not be able to properly recontextualize it.

The role of media companies in a networked society is to track and understand community needs and to help different communities with their information needs and to act as a mediator between communities. Understanding, following, and serving small and dynamic communities of practice in the best possible way has not been a focus of news publishing. However, news can be considered as boundary objects that media organizations can offer to communities of practice, and same rules for contextualization, decontextualization, and recontextualization apply to news stories, as they create links between communities.

4.3.3 Situated expertise and knowledge creation

In interaction people develop and then rely on *shared understanding of the domain of conversation*. This provides sufficient background knowledge to enable the participants to communicate with a minimum conscious effort. While talking, people are generally not aware of the structure of the conversation, because their attention is focused on the purpose and the content of the messages, and on the information that passes between the participants in conversation. This is true under normal circumstances, but when misunderstanding, or some other breakdown, occurs, people's knowledge of linguistic and conversational

structures is brought explicitly into focus to repair the breakdown. Such phenomena play a central role in human understanding and action. Therefore, designers should aim to achieve *structural coupling* between machines and their human users. They should not only design the artifacts themselves, but also the domains in which their actions are generated and interpreted [Winograd and Flores, 1986].

Can background knowledge be represented as an enumerated set of facts about context? Suchman [1987] argues that it is impossible to enumerate a possibly infinite set of contextual knowledge. In the course of interaction people rely on implicit information about the physical and social context that determines how they interpret what is said. This information is generally not available to the computer, and thus any elaboration of background assumptions using the formalisms of explicit knowledge representation is, according to Suchman, arbitrary and ad hoc. The formal models that underpin the machine operation do not as a rule correspond to users' purposeful activities – they are not properly relevant to real-world situations. This is largely because formal models are insensitive to the particular circumstances of action and interaction [Suchman, 1987].

The main design challenge in computer-mediated communication systems is to engineer resources to enable the computer to overcome its inherent limitations in interaction with people. In particular, the designers need to understand [Devlin and Rosenberg, 1996],

- how to extend the access of the machine to the actions and circumstances of the user to improve mutual intelligibility,
- how to make clear to the user what the machine limitations are, and
- how to find ways of compensating for the machine's lack of access to the user's situation with computationally available alternatives.

The tools should help community members to enhance information relevance and value of communication with system adaptation to user's preferences and needs. Community members should be able to link and reorganize information in order to build and update useful structuring of the information space. This serves as a motivation for the discussion on user and community modeling in the chapter 6.1.

Social creativity is impossible in communities in which most of their members regard themselves as consumers. Individuals acquire a new mindset when they become active researchers, constructors, and communicators of collaborative knowledge creation. Empowering individuals with tools is grounded in the belief that humans (although not all of them, not at all times, not in all contexts) want to be and act as contributors and designers [Fischer, 1998].

Computational environments support the needs of individual users, and social communities such as communities of practice, groups, and organizations. Individuality makes a difference, and organizations get their strength to a large extent from the creativity and engagement of their individuals. Organization amplifies individually created knowledge and crystallizes it as part of the knowledge network of the organization [Nonaka and Takeuchi, 1995]. Knowledge creation is the result of innovation, creativity, and

communication in individual members and among co-members of the community that evolve among individuals seeking to solve a particular problem or pursuing other commonly held objectives. There is no individual creativity without social support and there is no social creativity without creative contributions of individuals, thus there is a *mutually constitutive relationship of the individual and the social* [Brown and Duguid, 1992]. How could a media company better support skilled domain workers who are neither novices nor naive users, but who are interested in their work and who see the computer mediated communication as a means rather than as an end? How should we create co-evolutionary environments, in which users change, because they learn, and in which news systems change, because users become co-developers and designers of communication systems?

4.4 Implications for customized news

Stories that form the news are a potential source for learning about the world. Story-based models of learning emphasize the role of schemata and conceptual models that people use when they try to make sense of their environment. The important point from the perspective of customization is that stories are better understood if they can be related with something that has been learned before. If a news service can contextualize the story individually, the story will be more likely remembered.

The metadata-based approach for customized news services is motivated by research that states that domain models, and especially hierarchical categorizations are important in making the connections between the concept space (individual's models of the world) and news stories. Although these conceptual models are fluid and negotiated in social interaction, they form a basis for making sense about the news events. Therefore, as metadata for news is created, and customized services are built, some domain model is needed in explicit form. A customized news service utilizes these conceptual models to tailor the news package to different characteristics of individuals and communities.

The social character of news can be supported with customized and dynamic media. Small communities are typically places where the meaning of news is negotiated in a social setting. Customized news services can serve small communities with their specific needs, and by making the news better understandable by putting the stories into community context.

5 Customizing news

Digital media products have great level of flexibility. After the content has been created, news packages are built dynamically from content databases with a marginal cost to the news producer. Automated news system can select the stories for users according to their areas of interests, and level of expertise. An online journalist may enhance the news content with hyperlinks to original documents such as press releases and annotated reference material, links to other sites with information, and a selection of material in the news archives, as seen with CNN's Kosovo crisis theme site in chapter 2.2.

The following analysis (Figure 16) maps the different approaches of news content customization into two dimensions: *focusing vs. augmenting*, and *individual vs. community*. The first dimension describes the function of customization (what does the customizing system do), and the second dimension specifies the users of customized service (who the customization is for). Also in many cases the second dimension defines who gives explicit directions for customization system.



Figure 16. Dimensions of customized news content

Of these four quadrants, *The Daily Me* has received most attention in the research and business communities. The focus of this thesis is on the three other, less explored, quadrants of *The Daily Us*, *The World and Me*, and *The World and Us*. The following analysis includes a description of each quadrant and some interesting systems that fit in each.

5.1 Focusing for individuals

Term *information focusing* includes filtering, recommending, and prioritizing of information. Personalization has been one of the key elements of many new Internet-based information systems. The idea of personalized software agent that takes care of individual's varying information needs on user's behalf has attracted a lot of interest. Many commercial news providers on the Web have added personalization onto their services [Keane and Mooradian, 1998; Marriott, 1998]. In an empirical study examining uses and gratifications of online newspapers from the perspective of the audience, sites with personalization features were valued more highly than those without [Mings, 1998].

5.1.1 Filtering

Information filtering emphasizes the function of leaving out unnecessary pieces of information from constantly available and evolving information streams. Customers can select their areas of interests and be very specific about their needs, and automated news system will select those contents as they appear. Information filtering is not new, as first related projects and systems date back to the 1960's. At that time many SDI (Selective Dissemination of Information) systems were built and implemented. However, the large amounts and easy availability of information on the Internet, and increasing number of non-expert computer users have made information filtering available for everybody, as was seen with My Yahoo! in chapter 2.1.

The goal of information filtering is to select only those information objects that are relevant to user's interests, and therefore information filtering is closely related to information retrieval [Belkin and Croft, 1992]. Information retrieval involves short-term and immediate needs, whereas information filtering is based on repeated interactions over multiple sessions. Filtering systems typically learn over time to adapt to their users. This is desirable, since it is difficult to define an efficient search for topics of known interest, and it is even harder to know what the search should be about when the interests are not clearly specified. This is one of the main reasons why people follow news, as they trust the editors to provide those pieces of information they need.

Key challenges in information filtering are in describing the content metadata, in modeling the changing short-term and long-term information needs of the users, and in providing a scalable matching architecture for information brokering. The ideal "personalized news agent" has domain-specific knowledge, descriptive information (i.e. metadata) on contents to be filtered, and is able to relate these metadata structures to the user's interests. These services have moved from laboratories to commercial systems and there is a multitude of systems on the Web that provide information filtering for individuals at different levels of sophistication.

As a result of the *Information Lens* project, Malone et al. [1987] introduced three approaches to information filtering: *cognitive*, *social*, and *economic*.

Cognitive (or content-based) filtering

Cognitive filtering is based on the characteristics extracted from the available objects. Cognitive filtering is often also referred to as *content-based* filtering. Representations of the contents of information objects are compared with a representation of the information needs of potential recipients and then using these representations to intelligently match the objects to receivers. My Yahoo! (chapter 2.1) is an example of cognitive filtering system.

fishWrap [Chesnais et al., 1995] was one of the first prototypes of personalized newspapers using profiles of individual members of MIT community. fishWrap selects news from areas of interests included in the user profile. Each news story is accompanied with a ZIP code to build the hometown news section for each reader. Topical selections are made based on a categorization of interesting topics maintained by fishWrap administrators. In a larger commercial scale this model of filtering is used by information brokering companies, such as *NewsEDGE*⁵, that provide a service of high quality categorization of information as well as delivery of messages to companies based on organizational and personal needs. NewsEDGE is one of the largest infomediaries in the news business. They have filtering services for both individuals and corporate community customers. NewsEDGE relies on semiautomatic metadata creation for vast quantities of news items. This means that although most of metadata is machine-generated, human editorial staff assures the quality of metadata.

Content-based filtering is also used in "push" services, like *Pointcast*⁶. The content offerings are divided into channels (Business, Computing, Sports, Weather, etc.) that the user can subscribe and tailor to their liking with some parameters. The content resides in central databases from which packages are selected for individuals through the channels they have selected. After a period of fast growth, companies like Pointcast have stumbled. This is partly due to technical reasons as push services consumed too much network bandwidth and local resources (i.e. disk space). More importantly, the customers on the net rejected the idea of personalized content constantly sent to local terminals. The customers have preferred "pulling" or "browsing" personalized content, which can be combined with carefully designed alerting services and periodical summaries typically sent over email.

Filtering is valuable for individuals and business organizations that want to react quickly to news that have an effect on the market. Fast access to accurate information filtering transfers directly into money. A good example is the *Wall Street Journal Interactive Edition*⁷, which started a personalized section called "Personal Journal" in 1996. The system had initially a pre-defined limited list of categories, or a filter that

⁵ http://www.newsedge.com/

⁶ http://www.pointcast.com/

⁷ http://www.wsj.com/

consisted of free-text phrases entered by the customer. This was added by giving the user the option to create multiple personal "folders" each having a fixed category selections and key phrases.

The editorial content in *Wall Street Journal* is tagged during the editorial process. *WSJ Interactive Edition* staff members review the stories, and add more metadata to them, typically at the enhanced level of conceptual granularity to improve the category-based filtering.

An important lesson learned from the *WSJ Interactive Edition* is that customers don't want personalized information services to the exclusion of seeing the whole picture. They do value personalization *in addition to* the "broadcast" version of the site [Seybold, 1998].

Social information filtering

The term *social information filtering* has been used as synonymous to *collaborative filtering* systems (developers of the *Tapestry* system [Goldberg et al., 1992]) and *recommender systems*. Recommender system is a general term for systems that recommend items based on user evaluations. On the other hand most of these systems are not really collaborative, since people do not know each other directly, and there is no real collaboration between the members of the community of recommenders and recipients of recommendations [Resnick and Varian, 1997]. Social information filtering techniques group users into neighborhoods based on similarities between user profiles, and use these neighborhoods to recommend new items to similar users [Shardanand and Maes, 1995].

The communities in these systems are typically *ephemeral* (they do not exist for communications between community members, as users are typically not aware of the identity of other similar users) and *computational* (the community, or neighborhood of users, exists only for the purpose of making similarity calculations between individuals). Some systems have combined the recommending function with matchmaking between people with similar interests.

Hill and Terveen [1996] have explored how to gather Web recommendations based on their use in NetNews articles. Their system called PHOAKS (People Helping One Another Know Stuff) uses social information filtering to analyze people's recommendations in newsgroups. Essentially you can search on any topic, and the most frequently mentioned sites will rise to the top of the results list.

Content-based filtering and social information filtering mechanisms have also been successfully combined [Balabanovic, 1998; Pazzani, 1999].

Economic filtering

Economic filtering systems select data objects based on computations of cost-versus-value to the user. Malone et al. [1987] introduced economic filtering as a generalization of Peter Denning's idea of "threshold reception" [Denning, 1982]. In Denning's scenario, electronic mailboxes have threshold values that message sender needs to pay to get the message read by the recipient. Economic filtering criteria are based on a scarce resource that the items to be filtered are to consume. This scarce resource can be the *currency* available for paying for intellectual property and annotations that form the news content, the *time* required to digest the information, or the *space and computational capacity* requirements of available hardware and network connection.

The number of implemented filtering systems that uses market-based information filtering as the core of the filtering criteria is small. Amalthaea [Moukas and Maes, 1998] modeled the filtering task as a market of competing agents that internally use a bidding system to compete for getting through the filter.

Oard [1996] argues that price and other market-oriented metadata are just another form of annotations and that they can be used as criteria for more general content-based information filtering mechanisms. However, the term economic filtering has its merits as it emphasizes the cost calculation involved in information consumption in addition to the value calculation, as typically information filtering is based only on the estimated value of information to the user.

5.1.2 Prioritizing

In addition to filtering, news focusing includes prioritization of messages. Prioritization systems might show all information, but in a manner that highlights the most relevant information to the individual in a personalized way. Different prioritization mechanisms for tailored news content include sorting, highlighting, and re-sizing of news stories. Therefore, prioritization often includes some form of tailored presentation.

Customized prioritization can be combined with filtering, meaning that the filtered news will be presented in an order that reflects customer profile. Prioritization can also be used without filtering by essentially showing the same information content to everybody, but by simultaneously using different mechanism to customize the presentation. Newspapers use a well-established prioritization mechanism of headline size and amount of space given to a story. The Krakatoa Chronicle [Sakagami and Kamba, 1997] uses this newspaper metaphor on the computer screen to customize the layout of a personalized online newspaper. Another option is to highlight (for example, with different colors) the news page according to customer preferences, as though somebody had already gone through the newspaper and underlined the most relevant articles for individual users. This approach can circumvent some of the problems related to narrowing scope of interests and lack of serendipidity inherent in information filtering. Different models of customized annotation are largely unexplored, and in a sense are closer to augmentation of news as described in chapters 5.3 and 5.4. Important articles are Headlines move automatically shown at a higher position. from right to left.



The article's first sentence (summary) is shown when a mouse cursor is moved over the headline text.

New articles are shown brighter and larger, as well as moving faster than old articles, in order to create the illusion of a "close" object. The whole article (detail) is shown when the headline is clicked.

Figure 17. Anatagonomy/SS (picture from Sakagami et al. [1998])

Prioritizing can be adaptive. For example, the order of personalized sections in fishWrap reflects the usage patterns of the readers. The paper is organized in an adaptive fashion as topics of interest are prioritized higher than less interesting topics, and they are gradually rising up in the list of possible topics. This kind of adaptivity can also clash with the news customer's natural wish for routine and continuity.

Anatagonomy/SS (Figure 17) uses an approach that mixes content-based filtering with visualization of relative position, motion and perspective in conveying *information freshness* [Sakagami et al., 1998]. News articles are rendered as headlines that flow across the screen. Articles that best match the user's interest profile are at the top of the display. New articles are shown brighter and larger, as well as moving faster than old articles in the news flow, so they appear to be closer to the reader.

5.2 Focusing for communities

One of the possible negative implications of information filtering is the loss of a shared news experience across a community. Without a common base of reference, personalized news items cannot be shared, debated, and discussed like the common headline news of traditional media.

Malone et. al. [1987] argued that although users are often viewed as independent entities, they usually operate in a social context and are likely to share a significant portion of their interests with other users. This kind of community-based filtering is one of the focus areas of this thesis. How do people use news in a community setting to share information of urgency and interest? How can a news producer help communities in focusing on right bits of information?

fishWrap Page One is an attempt to solve this problem of splintering of community. fishWrap readers can, in addition to specifying the topics of interest or choosing to receive news from their home region, follow the interests of the fishWrap reader community. While browsing articles, readers can add meaningful and interesting articles to Page One. fishWrap presents the latest recommended articles on Page One in a ranked order reflecting the interests of the whole community. The readers vote the fishWrap Page One every day, so the community decides what's important and what is not [Chesnais et al., 1995].

Because communities create electronic artifacts that are negotiated explicitly amongst the group members, this information can be used in *community modeling*. Communities provide information about themselves in a more easily machine accessible form than individuals. Individuals rarely want to spend a lot of extra effort in defining themselves. The interests, roles and motivations of an individual can change quickly and can be very hard or impossible to detect by a computer. Community models are more stable and long-standing, and therefore possibly a better source for customization than individual models.

Knowledge management systems often aim to provide focused information to the business organization. InformationLens [Malone, 1987] is an example project that supports this kind of content-based or community-based filtering. Users can subscribe to, or be assigned to, real communities. The functionality of community-based information filtering systems rests on the assumption that the users know and trust some other users enough to accept their recommendations. When we get a recommendation from someone, the first reaction is often to question the personal relation to the recommender and how does it make me trust his/her judgment?

Community-based filtering has a team-oriented focus. The communities can have their specific interests that do not have to be the interest of each individual member of the community. An example may be a community inside a corporate organization that has jointly specified information needs as described in the *community profile*. Corporate Intranet with filtered information on topics of interest is an example of a focusing service for communities.

There is a lot of value in combining similarly minded people and giving them all an equal communications channel. However, this raises the problem of quality of the messages. There are often too many messages, of which only a handful is useful. GroupLens [Resnick *et al.*, 1994] allows people to rate Usenet articles, and recommendations are provided based on the similarity between the user and other users. PICS (Platform for Internet Content Selection) is a protocol specification for setting up rating Web pages [World Wide Web Consortium, 1997]. Internet mailing lists and newsgroups are good examples of content that give value to a specific community and are mostly created by the community members themselves [Resnick and Miller, 1996]. Starr et al. [1996] reported on "Do-I-Care"-agent (DICA), which allows users to collaboratively watch Web pages for interesting changes, and in which users can use others' learning results to augment their information gathering.

5.3 Augmentation for individuals

Information augmentation (IA) complements existing information acquisition, retrieval, distribution, and brokering tools. It is specifically designed to enhance understanding of content. With news stories, augmenting means adding explanations to existing articles, so augmentation is a way to provide context and background to news events. Augmenting complements focusing as the reader receives personally contextualized news. Customized augmentation increases the relevance of news by showing connections and comparisons to reader's history and current knowledge.

One important question with augmentation is to specify how much, and in what journalistic dimensions of news reporting (depth, breadth, distance), should the news story be augmented. To provide *depth* in customized augmentation, the story can provide historical and other kinds of analytic perspective to news, which are analyzed in a way that makes it easier to the customer to understand the reasons behind the news. The Kosovo augmentation example in chapter 2.2, significantly increased the depth of news coverage about the crisis. Historical perspective based on previous news is important, because often the meaning is uncovered gradually and understood when the whole series of stories of a news event has been published. To provide *breadth*, each news article can be accompanied with commentary from worldwide sources on the topic of question. This allows the customer to make comparisons and value judgments between different and often conflicting opinions on the news events. Another option is to include breadth automatically to the news stories by linking them with corresponding commentaries and related material. Perceived *distance* can be shortened by mechanisms of comparison between the news *event context* and the *individual and community context*.

Tailored views

Tailored views are adaptively created hyperdocuments, which are relevant to a particular individual or community context. The Galaxy of News system [Rennison, 1994; Rennison, 1995] organizes news stories by topic and presents the resulting topics in an interactive 3-D interface. *Part II: Publication 4* in this thesis introduces a system for financial news augmentation including tailored views that take into account customer's interests and previous experience, as well as community context of the customer.

Comparisons

Comparisons present information about a concept depending on the estimated level of user's knowledge of the concept and other related concepts. Comparison is an important and useful tool for understanding new concepts by relating that concept to other known concepts and also for helping user's misconceptions about the concepts in question. If another concept similar to the concept being presented is known, the user sees a comparative explanation, which stress similarities and differences between the current concept and the related one. News simulations based on comparison (what if the event would have happened in news

customer's own context) are potentially very efficient in increasing the potential of knowledge creation based on news events.

News story or news event similarity examination is based on document metadata about concepts in the news story. Once a concept is identified as potential for comparison with a concept in the focused news story, the system needs to determine which properties of both concepts should be used to liken and distinguish them. Two concepts are generally regarded as similar if they share several salient properties. There can also be criteria for selecting concepts that are dissimilar in an illuminating way. The comparison between concepts is created for comparing the most interesting similarities and dissimilarities. Both the news producer and the news customer can create the rules for comparison criteria.

Explanations

Explanations provide clarifications of concepts in the news material. This method can also attempt to hide the explanations that are not relevant to the user's level of knowledge about this concept. Some category of users can get explanations, which are specially prepared for this user group and will not be shown to other categories of users. This method is used in MetaDoc [Boyle and Encarnacion, 1994] where adaptive hypermedia presentation of instructive material was reported to reduce the time for learning and to improve the understanding of contents.

Pointers

Pointers refer to augmentation with related links. For example, *Wall Street Journal Interactive Edition* has the *Company Briefing Book* that is linked to each article where a company is mentioned and background information for the company is available. This company specific information is a combination of a financial overview, recent stock performance, company news, and recent press releases. The ability to drill down to get more detailed information about a topic that interests them was reported as critical to *WSJ Interactive Edition's* customer satisfaction [Seybold, 1998].

In a customized scenario, these pointers are provided automatically for each individual. The Remembrance Agent [Rhodes, 1996] suggests information relevant to the current usage context and acts as an augmented associative memory. For example, the word-processor version of the Remembrance Agent continuously updates a list of documents relevant to what's being typed or read in a word processing buffer.

Buffers Files Tools Edit Search Help		
email and starts editing a file, the RA automatically changes it	٦	
recommendations accordingly. These suggestions are presented in the form		
of one-line summaries at the bottom of the screen. Here they can be easily		
ignored, or the full text of the suggestion can be brought up with a single		
keystroke.		
Most applications for augmenting human memory, e.g. those developed by		
(Jones 1986) and (Lamming & Flynn 1994), have concentrated on		
Emacs: remembrance-agent.txt 11:44am 0.05 (Text Fill)L27 9%		
1 0.41 Felice Napolitan 24 Jan 96 Remembrance Agent talk/discussion	٦	
2 0.33 Brad Rhodes 25 Jan 96 Remembrance Agent available for Be		
3 0.31 Sumit Basu 14 Dec 95 Re: keystrokes		
4 0.16 fellowship testarne Oct 23 1995orientation forms in o		
remem-display		

Figure 18. The Remembrance Agent suggests relevant documents based on current document context

Autonomy, Inc. has commercialized the idea of the Remembrance Agent in their ActiveKnowledge system (Figure 19). The letter being written in Microsoft Word, as well as the Web page being browsed in Internet Explorer, are dynamically analyzed and compared against a database of local documents and external data sources. Pointers to related information are constantly displayed in the ActiveKnowledge window.



Figure 19. Autonomy ActiveKnowledge™ recommends related pointers

Berghel [1999b] uses the term *multiway lookahead* to mean the notion of computing several related items the user might want to see next and displaying them simultaneously. The customer can continue to read one or more of the pre-computed items without the mental overhead of thinking about requesting what to see next. This idea has been used in Letizia [Lieberman, 1995], which helps the user to concentrate on the additional material related to their recently visited pages. By making available to users what they will want to read next, without their having to explicitly click for it or otherwise request it, the user interface will demand significantly less attention and manipulation. Similar system called *Alexa* is bundled as the *"What's related"* feature in Netscape Communicator. It uses information about the Web-surfing paths taken by Web users to make recommendations about related Web sites. However, the augmentations in Alexa are not customized as users remain anonymous and everyone receives the same set of recommendations for a page.

5.4 Augmentation for communities

Community-based augmentation means contextualizing news to add relevance of news stories with additional community-specific information. A community profile permits tailoring news to all members of the community. The augmentation mechanisms for tailored views, comparisons, explanations, and pointers can be used also for community-based augmentation. Why is this news piece important for us? What would this news mean if it would happen in our community context?

Sara Elo's PLUM (*Peace, Love, and Understanding Machine*) [Elo, 1995] extends a news story to improve the reader's experience and understanding. Working with disaster articles, PLUM provides the reader with analogies to their local community to help them understand the significance of things like death tolls, financial losses, and property damage. It also helps put the disaster into a historical context for both the affected region and the reader's region.

PLUM is a text augmentation system for geographical communities, which uses fairly simple rules to make abstract data that appear in the story to be more relevant for the user. The user model in PLUM is simple: the system only knows the hometown of the user. However, PLUM has a rich knowledge base of statistical information on these geographical communities, and tools to make illustrative comparisons between them. In many cases this form of *community modeling* is preferable instead of individual modeling. For PLUM, there are many sources available where community profile information can be reached in a well-structured form, as Elo writes [1996]:

A computer is not able to record completely our personal experiences, unless is has detailed and up-to-date information about each reader. Such information is hard to acquire and maintain. It is easier with publicly available information on a geographic community. Information about demographics, weather history and geography of a city evolves more slowly than information about an individual. The shared information about the community does not need to be secured for privacy concerns. It's easier to contextualize news to person's community than to an individual in a community.



Figure 20. PLUM map augmentation [Elo, 1996]

One example of PLUM's augmentations is a map overlay that compares the disaster struck land area with the area of the user's home community. If the reader lives in the Boston area, an area under a flood in China is compared with a map of Greater Boston area (Figure 20).

Footprints [Wexelblat, 1999] takes an access-based approach to community augmentation, as the visitors to a Web site should leave their "footprints" behind when visiting a link and traveling a path through the document space. These paths accumulate over time and show the most heavily, and less frequently, traveled areas, and areas where many people click to but then disappear from the site. Any visitor to the site can see the footprints and visitors can remain anonymous. *SenseMaker* [Baldonado and Winograd, 1997] has a similar idea than Footprints, although the implementation is quite different. SenseMaker supports collaboration on information-exploration tasks. It helps communities of information seeking users to benefit from the different structures created and paths taken by members of their community in performing similar information-exploration tasks.

5.5 Summary of findings

Many systems, both research-oriented and commercial, have been developed for information filtering using various levels of sophistication in adaptivity, prioritization, and delivery mechanisms. However, information filtering is not sufficient to cater all the needs of news customers, but it is an interesting value-added service:

Only in very few and specific cases is a 100% automated personalization useful or required (e.g. for specific stock charts, topics or regions). For media companies, personalization makes sense as an add-on, embedded into an existing editorial framework of general or related information [ACHIEVE, 1998].

Above mentioned personalization means information filtering. Other forms of customized services, especially news prioritization, news augmentation, and community orientation, have been less studied and commercial implementations are few.

Journalists prioritize and filter incoming information and transcribe it into a relevant form for their reader community. However, in traditional media the information is tailored to meet the knowledge-level of the little-below average readers of the potential customer community, as it is not possible to tailor the information to meet the needs of individuals or smaller communities. By combining the journalistic work of high-quality editing and fact-checking with automatic augmentation the news can become more relevant and understandable to the task at hand, thereby reducing the information overload and decontextualization problem introduced in chapter 3. The systems for customized augmentation for individuals and communities are discussed more in the chapters to follow.

In current commercial news services, community-oriented customization typically consists of virtual message boards and chat areas where people can talk about the current topics. The possibilities of community-oriented customization are much broader, as illustrated in this chapter.

6 Implementing news customization

This chapter introduces the main challenges in customer modeling and content modeling that are both prerequisites for customized news services. These modeling mechanisms served as implementation principles for example systems that are also introduced in this chapter.

6.1 Customer modeling

In the context of customized news services, customer modeling should be used for

- reducing the amount of unwanted information,
- supporting customers at the conceptual level in their information needs,
- closing knowledge gaps between the customer and the content presented by the system, and
- providing customized visualization of the information space.

Rich [1983] classifies analytical user modeling into three dimensions. A slightly modified version of this classification to fit the terminology used in this work is shown in Figure 21.



Figure 21. Three dimensions of customer modeling, adapted from [Rich, 1983]

The first dimension is the *target* for customer modeling, which can be either an *individual* user, or many users that belong to a *community*. Rich [1983] used the term *canonical* user model to describe user models that represent multiple individuals sharing similar characteristics. The second dimension is the *source* of user model information. The users themselves define *explicit* models manually, whereas *implicit* models are built automatically through monitoring user behavior and unobtrusively acquiring other information

about the user. Third dimension is *duration* of the validity of items in the user model. *Short-term* modeling focuses on building up a user model during a single session and takes into account the current task context of the user, and *long-term* modeling concentrates on information that changes more slowly over a series of sessions, and that can be monitored over a longer duration between sessions. The information gathered in short-term modeling can be used for building and refining long-term models of the users.

6.1.1 User modeling

The user model is computer-accessible presentation of information about an individual regarding specified domains of use. The user model consists of some or all of the following:

- data *explicitly* given by the user for the purposes of modeling,
- *observations* of user's behavior,
- *deductive beliefs* (inferred data about the user), and
- *stereotypical beliefs* derived from available user stereotypes.

Customized hypermedia systems typically apply user modeling to different *goals, expertise, interests, preferences, background,* and *access history* of individuals [Brusilovsky, 1996].

Goals

Customers have different goals when accessing an information system. There are many reasons why people follow news and what explicit or implicit goals they want to fulfill by doing that. However, inferring the user's goal is a very difficult problem in current Web-based hypermedia systems with very limited possibilities for guessing user's intentions [Stadnyk and Kass, 1992; Höök et al., 1996]. The user can specify the goal explicitly, but this does not apply well to casual news usage, which is rarely a well-defined goal-oriented activity.

Expertise

The level of expertise of the customer regarding the content domain is an important dimension in customization, as user's information needs change as a result of interaction with information. Explicit models of the user's expertise have been used in educational adaptive *courseware* [Pilar da Silva, 1998], but expertise modeling has generally not been applied to news content. However, for the purposes of personalized news augmentation, it is useful to gather a model of user expertise on different domain concepts. The system may assume that the customer reads all the content displayed but that their knowledge of concepts diminishes over time. Also, an experienced user can get detailed and deep information while a novice can receive additional explanations.

In adaptive tutoring systems, user's knowledge of a subject is often represented as an *overlay model* [Greer and McCalla, 1993]. The overlay model is based on the structural model, or ontology, of the subject domain. The structural domain model is represented as a hierarchy or a network of domain concepts. The overlay model represents an individual user's knowledge of the subject as an "overlay" of the domain

model. For each concept in the domain model, the overlay model stores some value, which is an estimation of the user's level of knowledge about this concept. This is represented as a set of *concept -value* pairs with one pair for each domain concept. The value can be binary (*known - not known*), a qualitative measure (*good -average -poor*), or a *probability* that the user knows the concept.



Figure 22. Levels of expertise in a hierarchical conceptual model

A simple way to represent a user model is by means of a set of tuples $\langle u, c, v \rangle$ where u is the monitored user, c is a concept, and v is a value which indicates the level of relationship between the user and the concept. This relationship denotes that v represents can be user's expertise of c. The data types used to indicate these relationships could be *Boolean values* (true and false), *named values* (novice, intermediate, expert), and *numerical values* (between 0 and 100).

Interests

Customers have different short-term and long-term interests, which are modeled in the system. The interests need to be specified according to a representation mechanism that is useful for the adaptive systems. In the news domain, the interests are typically described as a weighted list of keywords. This mechanism of capturing information needs uses "keyword – value" pairs, which associate a weight with a given keyword (*"stock exchange"* – 0.8, *"motor sports"* – 0.35), and by combining keywords with query operators (Boolean, fuzzy matching or positional). Another approach is to specify the interests as overlay models, similarly to expertise modeling, and to assign weights based on domain ontology. The domain ontology used for describing the interests and expertise in the customer models is also used for describing the document metadata. This mechanism is used in the SmartPush project [Savia et al., 1998].

Tracking individual's interests is a challenging task for multiple reasons:

• Personal interests shift over time. People are very interested in earthquake information just after a big earthquake, but this interest gradually decreases over time.

- People cannot specify their interests, as they are often unconscious.
- Casual user's interests are difficult to specify and they are not willing to spend a much effort teaching their information agent their preferences.
- The domain ontologies that are used to describe personal interests change over time (see *Part II: Publication 5*).

Dynamic user profiles that learn over time what the user is interested in and monitoring the interaction of the user with the application. Learning algorithms make the profile adapt more closely to the viewer's habits. Learning system that monitors user's actions should also consider temporal dependency of user interests. Mostafa *et al.* [1997] show that a news filtering system that detects shifts in interests significantly improves filtering results.

Background

Customer's background means all the information, which is relevant for news customization, but is not directly linked to the news content, such as demographics. Traditional customer analysis and market segmentation is typically based on demographic background information about the users. For example, user's gender, age, education, type and field of profession can be used in making stereotypical predictions on the user's expertise and interests. The background information is useful when assigning a user an initial stereotypical profile. It is also useful when analyzing the usage of the service.

The system cannot learn these background parameters automatically, so they need to be explicitly asked from the customer. Also the demographic information changes over time (person's hometown, for example) so the explicit background information is error-prone.

Preferences

Preferences can include customized defaults for presentation layouts and typefaces, update frequency, alert mechanisms, and different delivery channels for messages. The customer usually provides these preferences explicitly, although some of them can be learned through monitoring usage.

Access history

Building a temporal model of user's behavior helps in predicting future behavior and in adapting to changes. The model can include tracking of repeating patterns of activity, frequency analysis, and long-term trend analysis. In addition to the statistical modeling, previous user history can directly be employed in system customization. For example, in electronic commerce the user can get recommendations for future purchases based on previous transactions or other events of logged activity. The server logs of the Web usage are adequate for this purpose as they act as a resource of user data for traveled document paths, time, and frequency of visits. The news customization system can store information about what the user has previously seen.
Stereotypes

A *stereotype* is the formal description of a prototypical user of a given kind [Rich, 1983]. Stereotype-based systems try to classify the users into predefined classes to represent the typical user categories. Stereotype is a collection of aspects of a user that often appears together. The stereotypical knowledge can be definite (necessarily true) or default descriptions about the user of a certain type.

The stereotypes are typically application dependent, and set up by the system developer beforehand, thus requiring the experience and bias of the system designer to recognize the most prominent stereotypes. The system can also be thought to recognize and extract stereotypical users automatically. If the producer predefines a collection of these stereotypes, a new user may start using a customized system with the stereotypical values as an initial user model.

Initial interview, typically in a form of questionnaire, allows for acquisition of primary assumptions about the user and is therefore a valuable source for assigning the user to a stereotype. User's relation to a stereotype model can be represented as a set of pairs of *stereotype -val ue*, where the value can be Boolean (*belongs – does not belong*), or a probabilistic value that the user belongs to the group of stereotype users. When a user is associated with a stereotype, all relations to hierarchical domain concepts that apply to the stereotype are activated.

User model acquisition

The user modeling system can acquire information either explicitly from the user or implicitly deducing the information from available usage data with machine learning techniques. Ideally, a user modeling system should not depend on users volunteering to give information about them to the system. Systems using implicit techniques observe the user and build the user model based on these observations. The learning data can be derived from the metadata about the media objects, and from observations about the user's behavior.

Explicit profile provision is the simplest method to gather data for the user profile, especially on background and interests. However, it puts the burden of *evaluation* to the user, and thus they have a greater mental load than they would with just reading documents and using a system. Explicit provision of profile information to the system adds to functionality that the user has to go through and learn. This happens typically at the stage of usage history when the user is not at all familiar with the system.

Interaction through the Web browser interface provides limited information for customer modeling [Boyle and Encarnacion, 1994; Kay and Kummerfield, 1996; Vassileva, 1996]. HTTP servers keep a detailed log of customer's patterns of user navigation, but it is hard to update the customer model using only this information. Rich metadata about the visited hypermedia nodes is an important addition for building and refining the customer models over time.

Indirect extraction of user preferences on the basis of the time spent reading a document is intuitively a reasonable data source for profiling interests because people tend to spend more time reading interesting articles than uninteresting ones. However, there are many problems with this approach. We are often subjected to interruptions while using a news service, so time spent per article should be used only as an implicit indicator of negative feedback (user selected the document, but backed away from it quickly). This approach is also problematic, since a short news story can be understood rapidly, and sometimes just seeing the title is sufficient for grasping the essence of the story. Often people glance through a story quickly, and decide to come back to it later when they have more time.

In certain environments user preferences can be acquired by monitoring user interface function calls, such as window scrolling. Sakagami and Kamba [1997] applied this method in a personalized newspaper, and they found that it is more useful as a personalization resource than strictly usage time-based monitoring, but the scrolling activity is useful only as an implicit indicator of positive feedback of user's interest.

Vassileva [1996] also argues that user needs are often imprecise, ambiguous, or unknown, particularly when dealing with novice users. Acquired profile information may be logically correct or have some degree of plausibility for being correct. Introducing uncertainty about the knowledge in the user model has great implications for the user modeling system.

Figure 23 illustrates one possible structure of a user modeling system, which is based on models presented by Huang et al. [1991] and Greer [1993]. The system must be able to "unlearn" the profile as user's needs change over time. This is the task of a *belief revision system*.



Figure 23. Modular structure of a user modeling system

Another question intercepting user activity is whether the user is allowed insight into, and maybe the possibility to change information in the user model. If this is desirable, a *model control system* is needed to

allow user access to review and change the contents of the user model in a convenient way. The possibility to check one's own customer model and change it, if necessary, is very important for building user's trust [Maes, 1994; Höök et al., 1996]. At any point the user should be able to get some explanation for the presented customization. The user can give feedback to the system after receiving the explanation. Since the system's explanations correspond directly to specific concepts represented in the user model, this form of feedback allows for direct changes to the induced model.

There are always important privacy implications when information about the individuals is gathered, regardless the mechanism of user modeling acquisition. The principle of *informed consent*, as implemented in *Platform for Privacy Preferences (P3P)* specification [World Wide Web Consortium, 1999b], should be applied. The user should be aware of what is being collected and how the collected information is used. Equally important is the ability to verify the contents of the user model.

Dynamics of user expertise and interests

As indicated in the previous section, expertise and interest are most promising user model contents for customizing news for individuals. The goal is to allow a system to specialize its actions to account for what the user knows and needs. A newcomer to the system should possibly be treated differently from an experienced user. If the user can reasonably be expected to know something, he or she should not be told it again unless specific evidence arises indicating that it has been forgotten. Many other attributes can be thought of, including a *knowledge decay* value (how much the user has forgotten about this concept).

The system should not assume that the user's expertise and interests are static. Learning algorithms applied to this task should be capable of adjusting to the user's changing features quickly, even after a long preceding training period. Therefore, a hybrid of user's long-term and short-term expertise and interests is needed.

The expertise and interest areas are closely linked. In information filtering a typical task is to find information that match user's interests and this requires a user model that can be static and explicit, or the model might be more dynamic based on implicit feedback and machine learning.

An overlay model is represented as a tree of <concept, expertise, interest> values, or V = <c, p(e), p(i)> tuples, one for each domain concept. The values of p(e) and p(i) are probability estimates that the user has expertise or is interested in the concept c.

The child nodes in the model can have negative interest values, but the sum of interest value at the root is always 100%, which represents relative interest for one metadata dimension. If there is no indication on user's interest and expertise on the domain, the default values for root node are 100 for interest and 0 for expertise.



Figure 24. Levels of interest and expertise in a hierarchical conceptual model

How does the expertise model correlate with the interest model? There are cases when the user might be interested but does not have any experience ("interested novice"). The opposite is a person who has lost interest in a field that he or she once mastered ("disenchanted expert"). However, in general the interest value serves as a reasonable assumption for the expertise value.

6.1.2 Community modeling

Communities benefit from information services and repositories that can, through use, self-organize and adapt to community needs. Community modeling has been little studied for the purposes of customized and adaptive news systems. Especially, there are no studies on the aspects of community of practice from this perspective. This is one of the main focuses of this work and of implemented system prototypes for news augmentation.

The mass media has traditionally been able to tailor the media to reflect the world from the perspective of a local community. Internet introduces new models for modeling and serving communities. The community-oriented journalism is typically directed towards geographically defined communities of proximity. I have instead concentrated in this work on communities of practice and the possibilities for profiling them. What kind of data should be captured in a community profile? How are the community profiles represented?

The meaning of stereotype profile and community profile is not the same. Collecting a sum of individual profiles and calculating their average rarely accomplishes a comprehensive community profile. Doppelgänger [Orwant, 1996] introduces the idea of *community beliefs* that are changing continuously as their constituents change, and can be generated automatically. However, this does not include the nature of communities themselves having a profile that cannot be amalgamated from individual user models, as Doppelgänger always computes community models as weighted combinations of their member profiles. In-built heterogeneity of communities of practice, resulting from the individual diversity and specialization

of their members, is often considered as a major strength in a community. Each participant in a community of practice finds a unique place and gains unique identity. These identities become interlocked and articulated with one another through mutual engagement [Wenger, 1998a].

Example communities of practice

To concretize the possibilities and challenges of community-oriented customization, I have described here some possible communities of practice and their customized needs regarding news services.

Symphony orchestra

An example of a community of practice might be a symphony orchestra. Probably the main (and maybe only) joint topic of interest of all members of the orchestra is classical music. Any single player in the orchestra might specify his or her own instrument as their special interest area. A trombone player would like to receive all news about other trombonists and the evolving skill of playing his instrument. These are the individual interests of members.

The symphony orchestra has a history of previous ups and downs, appearances in the news media, reviews about earlier performances, etc. The orchestra has plenty of ex-members, and different conductors have conducted the orchestra in the past. The community of practice creates a bond between current and ex-members.

Often the interests of the community of practice are decided for the community or by the leaders of the community. In the case of the symphony orchestra, the individual artists cannot influence directly the tour schedule of the orchestra. However, all members of the community might want to see what are the important events in the places they are visiting together, before, during, and after the visit.

Product design team

Second example of a community of practice might be a business organization that has assigned a team of employees to work on a new product. The design effort spans a large part of the company across organizational boundaries, from R&D, marketing, industrial design, and technology experts. These individuals have a joint enterprise to create a leading product for a market niche that is heavily competed. The company has fresh innovations that should be changed into products. Although different parts of the organization have their own isolated interests and areas of expertise, it is fundamentally important to enable sharing of discussion and exchange of ideas in this new cross-departmental community of practice. The information that relates to the market targeted for the new product, continuing analysis of competing companies, benchmarking against existing products their problems, and consumer reactions to product prototypes should be collected as a service for the whole community of practice. These are the contents of the community profile.

Silver Stringers

Third example of a community of practice is the Silver Stringers introduced in chapter 3.3, who are a local community of senior citizens actively debating, writing stories, and running their own Web publication.

Community profile

Community profiling is not just about electronic meetings, i.e. virtual communities [Rheingold, 1993]. The number of *physical communities* greatly outnumbers strictly virtual communities, and it might be that this division between real and virtual is not be practical in the future as electronic communication becomes widely adapted as a communication model.

Often there are big historical, topical and geographical differences between communities and most of these can be formalized into a community profile. The community can have a set of important values and issues of consideration that can be captured in a community profile. What are the most important issues facing the community currently? How could the community life be improved? What topics are causing conflicts in a community? What are the long-term values considered important in a community?

In their study of learning communities, Woodruff et al. [1996] identified four community "glue" factors, without which a healthy community will not develop and thrive: *function*, *identity*, *discourse*, and *shared values and interests*. These glue factors in mapping the self-defined aspects of community profiles.

Function

Function is the publicly articulated statement about the goals or objectives of the community, the joint enterprise that the members have engaged in. Not all communities have explicitly specified plans and goals, and some communities are not able to come to an agreement on what these might be. It is rare, however, that a community of practice does not have any definable function.

In most business organizations both long-term and short-term plans are known (at least to some degree) to all members of the community. In this community context, these plans should be reflected against information sources outside the organization, such as financial news.

Members of organizations sometimes have differing (and multiple) goals, and conflict may be as important as cooperation in obtaining issue resolutions. Resulting debate is an important source for understanding the nature of the community. This is not easy, given that there are often hidden or conflicting goals and people typically resist concretely articulating their goals. On the other hand, people are good at resolving communicative and activity breakdowns [Winograd and Flores, 1986].

Identity

Community identity is constructed though an understanding of member participation. Without a shared identity groups will not have the emotional bond to maintain group cohesion.

Current members, ex-members, and selected important people outside the community belong to the sphere of people attached to the community of practice. This circle of related people is important in defining community identity. Other important questions for defining community identity might be: What are the historical landmarks for the community? What are places of importance in community's history?

Over time, communities of practice build internal terminology that can be incomprehensible to outsiders and newcomers. The usage of vocabulary is one of the key elements in creating boundaries for communities of practice [Wenger, 1998a]. This is also an important issue when considering the need for customer ontologies and ontology mappings (see *Part II: Publication 5*).

Discourse

One important mechanism in understanding the interests of a community is to follow the debates and controversial discussions that the members of the community participate in. Conflicting viewpoints on issues of debate can be captured in the community model. The role of a news service is to "feed" the conversation in the community and enable the members to check their viewpoints against available information sources. Thus, community modeling also consists of understanding and modeling the topics of debate and controversy in the community. Community discourse can be understood as the evolving script, which creates and sustains the individual within the group.

A single member of a community of practice creates knowledge for the whole community as community members create shared assets. Thus, community of practice has a memory - *a historical trail of its past* - consisting of shared knowledge. This community memory can be systematically collected and stored in an electronic form. For example, one of the most typical shared resource for a community of practice or a community of interest on the Web is a collection of Frequently Asked Questions (FAQ). FAQ gives instant access to most often raised issues in the community.

There are topics of debate that are central to a community of practice. Historically important topics for the community are also significant. These can include issues that have already been resolved, but might be of interest to other communities as a community may be willing to share their solution.

Community discourse can consist of "war stories" (important events in the history of the community), stories of past members and their current whereabouts and doings, lessons learned (for example in business settings), and recordings of the significant past for the community. Community of practice often has internal databases of shared events, written plans, meeting agendas, and Web sites. These resources can be used to deduct the topics of interest of the community, for example when setting up an initial community profile.

Shared values and interests

Values that are shared amongst community members create common ground for debate. The culture of the community consists of values, rules, history, language, etc. What does the community like to promote

through the activities it's engaged in? What principles and values are shared among the members? What are most important topics and issues of discussion?

Table 3 enlists the some possible contents of self-defined community profiles for example communities of practice introduced earlier.

	Symphony orchestra	Product design team	Silver Stringers
Function	To better master the art of playing together To promote national culture while touring abroad	To utilize diverse skills in the team to make a product that beats the competition To fulfill company's business plan or strategic mission	To foster the creativity, stories and ideas available to others To actively debate on locally important issues (both online and off-line)
Identity	Artists with emphasis on modern music Conducted with world-class composers and played with well- known soloists	Company history, and R&D departments role Previous projects, their successes and failures Used vocabulary	Related people in Melrose and in neighborhood areas Milano Senior Center, their meeting place Their collectively created web site
Discourse	Meetings and rehearsal sessions Perfecting the playing Discussion about composers, new work Connecting with other musicians Future program for the orchestra, future tours	Design team meetings and presentation, group discussions Expectations from the management Coordinated activity and shared practices	Regular meetings and discussion sessions at the senior center Current events in Melrose Community memories Health
Shared values and interests	High artistic quality To work well as a team	Knowing everything about the problem domain to the new product Keeping the design project in schedule Confidentiality of information	Passing on stories and knowledge Continuing active participation in the community Connecting with children

Table 3. The analysis of possible community profile contents

Statistical information

In addition to the self-defined information, the community profile can consist of large amounts of collected statistical data about a community. This profiling has traditionally been done about local communities by governments and other agencies that aggregate statistical and demographic information about cities, towns and neighborhoods. This information can be extremely detailed, and can be useful in community-based customization.

Identifying and maintaining profiles for communities of practice

How are community models organized and created? The community profile can be initiated and maintained by hand meaning that selected individuals inside the community are responsible for modeling their community. Explicit modeling using forms is adequate for this purpose. News producer collects these community models, provided that the communities are willing to submit them. Electronic documents that

the community members create and collect, for example in the form of a web site, can also be used in automatically building an initial community profile that can be completed and corrected by the community members.

It is not simple to recognize how communities of practice form, change, and disappear. Typically members sign up for a community, but flexible tools are also required that unobtrusively identify the communities. For example *Visual Who* [Donath, 1995] used electronic mailing lists and the activity patterns of their members to visualize the aspects of community participation, boundaries, and membership.

6.2 Combining individual and community profiles

The use of individual profiles can be effectively combined with community profiles. What kind of knowledge representation should be used for these profiles? How to combine social information filtering and community-based augmentation?

People can be automatically clustered into functional groups that share similar interests or are otherwise like-minded. *Dynamic communities* mean aggregations of people created automatically, with some form of similarities amongst their profiles. This has mechanism has been used in social information filtering.



Figure 25. Modeling the importance of different communities for an individual.

Each user belongs to many different communities simultaneously. Membership in a community is matter of degree, rather than binary, as each community is a "cloud" that includes some people more than others. Doppelgänger allows communities to be treated as individual user model, so the system can build a news environment for a community. A news edition can be a combination of "80% me" and "20% my community". The numeric way of presentation has some preferences over the symbolic way: it opens the

possibility to combine several user models and to accumulate a group user model [Kaplan et al., 1993; Mathé and Chen, 1996].

The interconnected nature of user models, community models, and stereotypes is depicted in Figure 26. The community C_n has a collective profile $P(C_n)$. The community might also have associated stereotypic profiles that consist of typical individual profiles amongst the community members. They can be used when assigning a new individual profile of new members.

User U_n has an individual profile $P(U_n)$. An individual belongs to communities in different ways. This user-community relationship $R(U_n, C_n)$ to different communities, can be an abstract notion of weight or importance of the profile for the individual. The relationship can also contain other things, role, history, trusted other members, etc. The relationship R between users and communities includes the "degree of membership". Is the community important: is it central to the user or is it peripheral? What is the role that the member has in the community? How active is the user in the community? What is the history of members participation in the community?



Figure 26. The relationships between user profiles, community profiles, and stereotype profiles

There are also more general stereotypes that do not relate specifically to any single existing community, such as stereotypes that are solely based on demographic data about users. The user can be assigned to initially inherit the features of these "super-stereotypes" (SS_n).

An example of community-based augmentation is news presentation for corporate organizations. Usually people who worked in teams share similar professional interests. Currently the individuals in these working groups inform each other about interesting events and that can have influence on their own professional field of expertise.

6.3 Content modeling

This chapter discusses the importance of structured metadata in the customization of media products. Metadata is needed for both information filtering and information augmentation.

Library science has for centuries searched for usable ways of describing and categorizing information. In the pre-digital era of news media, the use of metadata on content in a media company has traditionally been limited to the archived versions of the products of publications. At the time of archiving oftentimes specialized librarians in the media company have been responsible for categorizing and keyword tagging the material.

Metadata contains a representation of the meaning, content, organization, or purpose of data. Metadata also includes information about the location and characteristics of the data objects, and about the logical structure of the data objects and relationships between them. Main advantages of metadata for media companies are

- improved production processes for new content,
- improved searchability of content,
- tailored products for individuals and small communities, and
- **augmented content** enabled by rich metadata descriptions.

6.3.1 Ontologies for semantic metadata

Ontology provides a set of well-founded constructs that can be leveraged to build meaningful higher-level knowledge. Ontology is also more than a taxonomy or classification of terms. Although taxonomy contributes to the semantics of a term in a vocabulary, ontologies can include richer relationships between terms (Figure 27).



Figure 27. The news story with relations to structured domain ontology

High-quality metadata requires a significant investment from the news producer. Compared to traditional publication the effort for automated customization in publishing process is elevated, since information has first to be structured according to metadata dimensions, in order to be applied to specific user profiles (Figure 28).

The issue of ontology development for media environments is further discussed in *Part II: Publication 3* and *Part II: Publication 5*.



Figure 28. Customer model organized according to metadata dimensions

Automatic language analysis based on syntactic information in a text of the news story can be an important way to enhance news content. Tools for natural language understanding, knowledge representation and heuristics, are used for this purpose. Information Extraction (IE) is the primary mechanism for spotting various types of concepts, such as proper names or locations, in text documents. This enables the creation of metadata that helps in identifying news articles or other information that is for example related to a company, a person, or a geographic area. Information extraction and automatic classification of multimedia content can be used in semi-automatic creation of metadata (see *Part II: Publication 3* and *Part II: Publication 5*).

6.3.2 Adaptive and dynamically created hypermedia news

An *adaptive hypermedia system* guides the reader towards relevant, interesting information using *link adaptation*, and provides additional or alternative information to show relevant information so that the reader can understand the information better as it is presented using *content adaptation* [De Bra, 1999]. In adaptive hypermedia, an existing network of documents (i.e. hypermedia nodes and links between them) are adapted by providing flexible document content and flexible link structures to users according to their interest and expertise of the document-related concepts. The system monitors the user's behavior and adapts the presentation accordingly, as the evolution of the user's preferences and knowledge can be attempted to deduce from collected usage data. Sometimes the system may need questionnaires or tests to get a more accurate impression of the user's needs.

Dynamic hypermedia system creates both the hypermedia network and the documents within the network when the user requests them. The hypermedia network is therefore completely *virtual*, and the structure and content of documents are created on-the-fly by the system, based on information retrieval techniques from heterogeneous databases, automatic text generators using rule-based systems, and information visualization techniques for document spaces. This *non-prescriptiveness* means the ability to transform or traverse an information artifact in ways, which were not prescribed by the information producer [Berghel, 1999a]. Adaptive hypermedia and dynamically created virtual documents can thus be used for effective visualization of query results from a traditional Information Retrieval system as a hypermedia structure. This is the approach of *"Historical context visualizer"* described in *Part II: Publication 4* in this thesis.

Adaptivity in hypermedia is generally proposed as a means to interact with users having different needs, background knowledge, interaction style, and cognitive characteristics. It is also seen as a partial solution to information overflow and navigation through large information spaces and ordinary hypermedia.

6.3.3 Matching customer profiles and content profiles

Customization of the content involves selecting the content on the basis of the metadata and the customer model. This profile-based matching differs from traditional database queries, or information retrieval methods. An agent-based matching architecture for information filtering is discussed in detail in *Part II: Publication 3*, and the issues of metadata matching, using distance measures between hierarchical metadata structures are presented in Savia *et al.* [1998].

6.4 Example systems

6.4.1 Adaptive news filtering

One focus area of research at the Ota Online project at Helsinki University of Technology was to understand the technology behind personalization. An early prototype for an agent-based architecture for information filtering using Java and Knowledge Query and Manipulation Language (KQML) was developed. This work is discussed in *Part II: Publication 1*.

The SmartPush project at Helsinki University of Technology has worked on adaptive news filtering which is based on rich semantic metadata. Although the author has not participated full-time in this project, he co-authored a paper on SmartPush approach. *Part II: Publication 2* introduces metadata-based information filtering, and describes the agent-based system that was built to implement it.

6.4.2 News augmentation system

The editor in a media organization has traditionally had the job of integrating information into a coherent story. Since there are not enough journalists to create customized documents to every potential reader and not even to every potential community, the customizations have to be made computationally. What is the space of possible augmentations? Which features of the news augmentation system can differ for different users? What can be adapted in this computational system?

News augmentation system augments content according to the individual context and community context. These contexts are captured in (1) individual model, (2) community models, (3) currently focused story, and (4) previously seen documents for user's immediate context. The system can then incorporate into currently focused story links to related concepts or entities in the story domain. The metadata of currently focused story is used to compute object similarity with other potential data objects to be included as part of the augmentation.

News customer should concentrate on information absorption, not on manipulating the interface. Therefore it is important that the interface is unobtrusive and transparent, to avoid interfering with the goal of maximizing the efficiency of information transfer to the customer. Also the performance of the system is very important in interactive settings where information to be presented is transient and must be recomputed frequently. The whole point of information customization is to speed up the transfer of useful information to a customer.

When a news augmentation is created, the following requirements need to met:

- *narrative coherency*: the story should be augmented with components that fit into the narrative framework of the story,
- *narrative links*: the links provided from the customized document should enhance the understanding of the original story, the purpose of the links should not be ambiguous,

- *redundancy management*: same content should not be duplicated multiple times in the augmentations, similarly a model of previously seen content should be utilized, and
- *volume constraints*: the size of augmentations should not overwhelm the user.

One design goal in the news augmentations system is to provide a *conceptual navigation environment with a queryless interface* that guides the user through the document space. The complexity of queries behind the interface should be hidden from the users.

Structured documents and data repositories

Information producers' data sources are described through the producer profiles, which capture the metadata of the sources such as the content, category, query capability, and access rights of the sources. Conceptual model acts as a translation layer between the heterogeneous databases and the query that is based on current customer context and customer model. To access information in terms of concepts, their attributes and their semantic relationships are combined by translating queries on these concepts into dynamic database queries.

To provide accessing heterogeneous data sources, a news augmentation system must have enough information about the content and query capability of the different data sources. These are specified in *source capability profiles*. The source capability profiles can be part of the producer profile, if the news producer manages the source, or it can be part of the consumer profile, if the source is managed by the user or user community. This is the case in integrating news sources with internal data repositories of customer organizations. One of the most important commodities for news sources is credibility. The source of both news stories and their augmentations should always be clearly stated.

The main problems in providing quality access to augmentation information sources:

- *autonomy and heterogeneity:* The sources are independently designed and managed. Their data models are different and changing over time. New data sources or applications need to be added to an already heterogeneous mix, but existing data sources may also change the specifications of the data they provide, and consumers may change the requirements for the data they request. Also, the number of sources available online is constantly growing. Autonomous data sources are often interrelated and possibly replicated.
- incompleteness: Most data sources contain incomplete information.
- *limited query-facilities:* Many data sources are not full-featured database systems and can answer only a small set of queries over their data.

To deal with these problems, the news augmentation system should allow users to pose queries on the fly in the sense that users can issue queries without knowing the structure, location, or existence of requested data. It is not realistic to maintain a pre-defined integrated view of all data sources due to the dynamic nature of the data sources online and the different objectives of information consumers. The volume of augmentation should not be overwhelming. Therefore, *a layered model of augmentation* is needed. This can be done with efficient structuring of contextual information. The layered model does not necessarily mean that the more expertise a person has on a domain, the more he/she will receive contextual information. In some cases this might be just the opposite as experts receive only information that is new as novices get a broader spectrum of information (see Figure 29).

If the gathering of related material on current events becomes easier and more automatic, the journalistic occupation will shift to sorting, evaluating, and creating a layered package of related information for augmentation.



Figure 29. Layered model of augmentation

The eXtensible Markup Language (XML) allows the description of electronic documents by means of a Document Type Definition (DTD). XML is intended to be evolutive since it allows for the fusing of several DTDs. XML is an effort by Web community to provide the machine-readability and interoperability to Web resources [World Wide Web Consortium, 1998].

The Resource Description Framework (RDF) provides facilities related to our problem. RDF is a framework for specifying metadata that enables automated processing of WWW resources. RDF uses XML as a common syntax for the exchange and processing of metadata. By exploiting the features of XML, RDF imposes structure that provides for the unambiguous expression of semantics and, as such, enables consistent encoding, exchange, and machine processing of standardized metadata [Lassila and Swick, 1999].

XML has been used as a bridge between the news stories and heterogeneous databases. For this purpose the XMLNews DTD was used⁸. XMLNews is heavily borrowing on the News Interchange Text Format (NITF), defined by the International Press Telecommunications Council (IPTC) and the Newspaper

⁸ XMLNews is available at http://www.xmlnews.org/

Association of America (NAA). The metadata about the news item is split into a separate file according to XMLNews-Meta DTD that conforms to RDF.

Conceptually most of the news data is stored as frames, but the system uses a SQL database for physical storage. A frame is a collection of key/value pairs, which describe a particular object or concept [Haase, 1996]. Each key/value pair is a record in the database, which enables fast response times on searching. The database also includes the XML versions (both XMLNews and XMLNews-Meta) of the story data, which is used to transport data between modules.

System architecture for news augmentation

Autonomous agents are programs, which carry out tasks on the user's behalf. The agent is expected to exhibit autonomy (object identity and independent behavior), social ability (interaction with other agents), reactivity (perception of the environment and reacting to changes), and proactiveness (initiating action to pursue goals) [Jennings and Wooldridge, 1998].

An agent requires information on their owner's personal preferences and intentions. Several existing agentbased implementations have been implemented, which include mechanisms for obtaining user modeling information. In agent-based systems, the adoption of a shared ontology allows agents to interoperate and to retain a high degree of autonomy, flexibility and agility. Information agents can be highly adaptable yet are able to meaningfully communicate domain-specific knowledge. They do this by using only the basic terms and relationships defined in the ontology. It is the precise definitions of the basic terms that allow those terms to be combined to form meaningful higher-level knowledge.

Agents may need to access only descriptors of the multimedia items, not their actual contents. Information source, producer or other agent, provides logical addresses for the selected items, and the agent presents these to the user. The selected items will then be fetched from the information producers and presented by the user's multimedia presentation control module. The cost of information to be delivered can also be presented to the user, before the actual delivery is done.

The architecture of the system consists of *profiling, mediator, and augmentation agents*. Profiling agent monitors individual users and user communities. Mediator agent connects various types of data resources and combines data into a common format used in augmentation. Augmentation agent sends a query request to the mediator agent. There are many augmentation agents in the system for different types of augmentations. The mediator agent combines current profile data with content metadata to form the query, and sends the query to search engines of different heterogeneous *information sources*, for example a full-text news article database. The results of the query are mediated to the augmentation agent with relevant profile information, which presents the augmentation to the reader.

Agents can reside on a server hosted by news service provider, or they can be on the Intranet of the organization that has subscribed to the service. On the Intranet, the company's internal databases and information flows provide additional organizational context for news augmentation.

For the purposes of user expertise modeling, I assume that interest profile and expertise profile regarding news content are highly correlated, especially in the case of professional interests. The same interest profile that is used in *information filtering* (IF) to provide focused news coverage on the topical areas of user's particular interests, is also used as a basis for expertise profiles. If the user has expressed interest in certain concepts in the conceptual model of content hierarchy, then these interest levels can be used as an estimate of the users expertise levels. Users submit queries to the mediator agent in terms of concepts, their attributes, and their semantic relationships, that are selected in the augmentation interface. The query engine then uses the conceptual model to translate these SQL queries and full-text queries.

User profile contains explicit information, extracted information from user's personal documents and email, and implicit observations of the user's behavior. Interest can be used as an initial estimate for the expertise profile.

Community profile contains background information about the community, the membership information between communities and individual users, and the interest profiles of the community. Community and individual profiles are combined by the mediator agent, which gives higher priority to data in the individual's profile.



Figure 30. Augmentation system architecture

Content metadata is a key requirement for the augmentation service. News content is enriched with semantic metadata in the early stages of the content development cycle. This is done by the news editors or by special personnel in charge of adding the metadata. In addition, some content-related metadata can be automatically created. When content has a metadata representation that can be matched with the user

profile and the community profile, the job of the mediator agent becomes relatively easy. It only has to compare the descriptions of content metadata against the profile data, and make query decisions based on the combination of these descriptions. Augmentation agents use XML as a transport mechanism. The agent then creates a suitable kind of presentation. The overall architecture of the augmentation system is depicted in (Figure 30).

Case: Kauppalehti Online - Financial news in context

The financial news augmentation system provides mechanisms to analyze the relationships between information objects and builds a representation of the underlying structure of the entire body of information. This relational structure is used to construct a visual information space with which the user interacts to explore the contents. The system can adaptively customize the presentation to user's interests.

An example for this approach is shown in (Figure 31), which takes a dynamic approach to producing context using proactive and personalized conceptual information retrieval. This prototype system has been implemented for the Web service of a Finnish financial newspaper, Kauppalehti. It automatically creates a presentation of related articles and visualizes the relationships between the concepts in these articles. The system uses a "margin notes" approach in displaying additional information, rather than embedding them as direct links in the text, to allow the reader to ignore the augmentations until they are needed.



Figure 31. Financial news story with customized augmentations

"Time series" augmentation adds background to numbers appearing in the news. Whenever there is numeric information in a news story, a graphical depiction of the history of that particular numerical information can be attached. This is kind of background information is typical to financial news sources. A company can also integrate and link their internal databases with external databases that have been used to provide the financial information

This financial news augmentation application is described in more detail in Part II: Publication 4.

6.5 Summary and implications

Information is always a social commodity, which flows not only among individuals, but also between levels of scale or context. Technology can empower individuals, and communities, by supporting the cross-domain and cross-context sharing and use of information and knowledge. However, this requires well-modeled content streams and customer profiles that can be produced and maintained by media organizations. Community modeling has some significant advantages over individual user modeling: the contents of the model are negotiated in a social setting and are more likely to remain more static over time.

Some form of profile is necessary in order to build customized services. However, when modeling and adapting to human interests and expertise the user model can be very coarse and often inaccurate. This does not mean that these profiles cannot be helpful, especially when comparing this amount of computer accessible information on user's interests and expertise with a similar human-to-human teaching situation.

New tools allow media companies to serve the individual from a personal viewpoint. With the growing size, complexity and heterogeneity of the Web, it becomes virtually impossible to provide navigational tools that, as such, provide support for heterogeneity of needs and backgrounds of all customers. Therefore, dynamic news systems should have the ability to adapt to the needs of their individual users, and user communities. Adaptive systems try to learn about their users and anticipate their needs and desires. In the case of news content, the purposes of adaptation are tailoring news presentations, recommending relevant information through information filtering, and helping to learn about the news topics using news augmentation. Adaptation to customer needs happen in multiple dimensions: expertise and domain knowledge; interests and preferences; goals, beliefs, plans and skills; and physical and sensorial abilities. When user and community profiles are combined, priority is given to individual preferences.

Dynamic community profiles and community-centric augmentation have not been discussed much in the literature. There has been work in augmenting news content using static background data and static user models. This work combines dynamic profiling with constantly changing content and augmentation resources.

7 Models for publishing customized content

News organizations provide a service of packaging and presenting the world events in an easily digestible form. This can be considered as the *generic news product*. Levitt [1980] argues that all products should be considered as "problem-solving tools", and that the producer should always think how different customers have different problems. He argues that all products can be customized and made special, as all products and services are combinations of features and possibilities. Only the most basic parts of products are commodities. Unique additions to a product or service can be based on a customized approach. Different forms of electronic communication, possibly including the element of customization, fill the media gap that was introduced in chapter 1.

However, there are multiple possible models for the implementation of customized news services, as discussed in this chapter. The key questions when considering these models are:

- Where is the data? Is it at the producer, locally at the customer, or distributed across the network?
- Who keeps the customer models (producer, intermediate trusted party, or customer)?
- What is the role of individual and community customization?
- Where is the computational load?
- What is the level of privacy?
- How do the production methods change because of customization?

7.1 Broadcasting model

Mass communications has traditionally meant that everyone in an audience receives the same, identical product. This broadcasting approach (Figure 32) has advantages in rapidly spreading information to large amounts of customers, and in quickly building up brand recognition for the content media company is providing. Advertisers, who are typically the main source of income for these channels, choose a broadcasting channel based on the number of people reached, as well as the brand image and content of the publication. Segmented publications address smaller groups of people, typically based on interest or demographics of the audience. Advertisers are targeting these same market segments with their product offerings [ACHIEVE, 1998].

The broadcasting model, as such, does not really support customization. Segmented publications are still not customized, as everybody in the market segment receives the same content. However, the received message can be customized locally, using the computational capacity of receiving device. This *local customization* on the client-side is possible in all of the publishing models discussed. In the broadcasting model it is the only option.



Figure 32. Broadcasting model with local customization

Broadcast content flows continuously and it is up to the customer to fetch the right bits synchronously. Two basic scenarios exist for local customization [Dimitrova, 1999]:

- standardization of metadata enables the end device to do the filtering based on metadata about the content streams; or
- no standardized descriptions are available and the end device needs to analyze the contents locally, meaning that the algorithms should be robust and optimized for the consumer device.

In a personalized scenario, the data offered to users can be filtered from broadcast streams according to individual profile. The preferences in the profile can be explicitly provided or implicitly derived by watching users' actions. This personal profile becomes a query, which continuously compares profile against the broadcast stream. The selection remains transparent to consumers because the query derives from their personal profile. The user can specify category, channels, anchors, frequency of news update, location, and any additional keywords. The broadcasting model does not have a privacy problem if there is no direct uplink to the producers.

7.2 Individualized model

In the broadcasting approach, the media company may know some common characteristics of their public, but they cannot customize the content to particular customers. Customization makes possible different business models than the market segmentation approach used in broadcasting, as the customers can be served based on their individual profiles (Figure 33).

The information producer maintains customer models. The content is stored in databases and enriched with necessary metadata to enable customization. As the media company learns the interests and the usage patterns of their customers, it can use gathered profiles to individualize both the editorial content as well as the advertisements and other commercial content. News content is pulled from the content repositories as

the customers wants to access the personalized version of news. The producer can also push the content towards the individual according to the customer profile.



Figure 33. Personalized news content is delivered according to an individual profile

The producer (server) and the customer (client machine) share the computational load. However, the burden is typically on servers administrated by the producer. Individualized model is prone to scalability problems on the server, as customization requires frequent database access and potentially intensive computation for profile maintenance and content matching. On the other hand, the overloading of client's resources (CPU, storage space, network bandwidth) was the main reason for customer rejection of push-based individualized systems in 1997-98.

Privacy is a major issue in this model, since typically the producer maintains profile information about their customers on the server. The profile information can also reside at the client end, which provides more control on the access and usage of profile data to the customer.

7.3 Infomediary model

The infomediary model - or the *portal* model - for content customization is similar to the individualized model. The infomediary mediates the news content between the producer organizations and the customers. Infomediary aggregates the content from different sources and also hosts the customer profiles used in customization. This is the model in My Yahoo! introduced in chapter 2.1. As a result, the infomediary has the customer profiles and controls the flow of communications and the customer relationship. The infomediary can be a traditional media company or new players can take this role.

The advantage of the infomediary versus traditional media companies is its position as an independent and trusted third party between the content producers and the customers. Infomediary combines and offers content from multiple, and possibly competing, content providers in one customized package to customers.



Figure 34. The infomediary collects news and routes them to communities and individuals

Computational load is mainly at the infomediary that becomes a new center node in the delivery chain, although the infomediary often only offers pointers to actual content that still resides at the producer's server. Technically this requires common standards for content syndication, such as *Information Content Exchange (ICE)*, and shared ontologies for semantic metadata (see *Part II: Publication 5*).

The infomediary can compare the popularity and effectiveness of different products and different producers. It also becomes an attractive spot for advertisers if the largest numbers of traffic are passed through these infomediaries, who at the same time hold rich profile information on the customers. Traditional providers are not likely to give this position away easily.

The news producer's brand image is an important part of guaranteeing the quality of the content, and in promoting trust amongst the customers. The main advantage of current media companies in comparison with their competitors is the capability of producing good quality content and the trust gained from successfully doing so. The customer develops over time a special relation with a media brand. The infomediary needs to build a similar kind of trust relationship with its clients, especially as it collects sensitive profile information.

7.4 Community-oriented model

In the community-oriented model the producer targets entire communities with customized products. This requires mechanisms for modeling the differences between communities. Traditionally, local newspaper journalism to communities of proximity, as well as magazine publishing to communities of interests, have used this model. This work has concentrated on possibilities in modeling communities of practice.



Figure 35. The news producer provides content for all members of the communities

This model is identical to individualized model, except that it does not have as high requirements for scalability and privacy. Typically managing community models is not as problematic as individual user models. An example of this model is a news service that a whole company has subscribed to, and that it available to all members inside the organization.

7.5 Community-mediator model

Typically, the preferences of a whole community are not able to satisfy particular individual's needs. There can be dedicated members in the community, i.e. *community mediators*, that follow the specified information flows, or community-oriented content packages (as in previous model), and act as human *information brokers* in distributing relevant material to other members of the community. This has traditionally been the role of professional information-seeking intermediaries in an organization. Human community mediators combine the skills of professional journalist, publisher, scientist and a librarian with machine intelligence in order to filter information for users based on their interests [Höök et al., 1997].



Figure 36. The community mediator routes news to individuals

The community mediator needs to be equipped with an efficient integrated environment for searching for information, utilizing a variety of tools for searching, selecting, restructuring/rewriting, and annotating information. In the community mediator model, the usage of a human editor will allow users to trust the system at the level they trust the professional editor responsible for the service. Also the privacy problem is not as severe, as the community members trust the community mediators.

Professional journalists, especially with local journalism, in media companies have the role of community mediator, but this role is rather impersonal. One could envision media services, where the journalists would provide the service of dedicated community mediator. In fact, for business information in company settings this has already happened, as information brokering companies are already doing this at a certain level, as demonstrated by NewsEDGE.

7.6 Community-centric model

The community-centric model means that the members of the community recommend items of interest to each other. There are trusted networks of connections inside communities, and between different communities. This model emphasizes the role of news as boundary objects that provide bridges and tools for communication between communities. The producer can facilitate this by supporting communities in providing comparisons and matchmaking between different communities (Figure 37). This can be achieved by collecting models of customer communities.



Figure 37. Community-centric model of customization

The rapid increase in the number of Internet users has made possible conversations in which hundreds, even thousands, of people participate in many-to-many communications. The most obvious manifestations of this phenomenon are Usenet newsgroups that are hosted on hundreds of thousands of servers on the Internet. More lately this community discussion activity has been used as a basis for commercial services that archive the newsgroup contents and act as a bridge between these online communities, external information sources, and advertisers.⁹ This illustrates, that there is value in providing a service that connects communities of practice.

In the future, the journalists may take the role of moderators of horizontal communication among the audience [Heinonen, 1999]. As a result, the role of news producer becomes more like a service, and less like information feeder to the consumers:

In the news-as-a-service model the consumer of news is an active, engaged participant. This service model encourages two-way communication between traditional news providers and the consumer, and communication within communities built upon common interests. The news service model becomes a part of the social fabric within communities, a catalyst for creating communities of interest, and a means of facilitating community insight [Bender, 1996].

7.7 Summary of findings

The broadcasting model with local customization is challenged with various other mechanisms for customization. These models can be targeted towards individual customers or customer communities. The role of infomediaries is also a crucial point in defining the structure of customized services.

The location of individual and community profiles used in profiling is important both from the business perspective and the technical point of view. This also has great implications to the privacy of customers. Consumers will become conscious of the fact that the information they reveal about themselves has great value for businesses using it. Therefore they start demanding value from the businesses using it. The "coming battle of customer information" will not so much relate to privacy of the individual, but rather on the growing expectation of return value from the information given [Hagel and Rayport, 1997].

There are multiple options where the customer profile information can be located:

- at the user's end terminal or a physical removable storage medium (like a smart card),
- in the network "cloud" at the operator's smart message routing facilities,
- at the infomediary that the user accesses most often when looking for certain types of information,
- at the information producers, like traditional media companies, and
- at a third trusted party, like banks, insurance companies, government officials.

⁹ Deja.com is an example of such a service (http://www.deja.com/)

The current model is that the customer information is dispersed as fragments into all of these.

The infomediary can make a useful service as the aggregator of information resources. However, the producers are unwilling to give away their direct relationship to the customer. Therefore, the producers are often trying to hold the position of both content producer and content aggregator.

8 Conclusions

In the converging world of digital communications, the media industry has struggled in finding ways to provide services that add real value for their end customers – the readers, viewers, and users of electronic media products. Many traditional media companies have found that the new media environment requires different business approaches than those they have been used to in the mass communications environment. Only a small number of media producers have been able to make Web services that the clients would be willing to be pay for. Also, media companies know how to build products that attract large audiences. The attention of these audiences is then sold to advertisers, who are looking for potential market segments for their product offerings. New forms of value-added customized publishing services are based on a long-lasting relationship between media companies and information- and entertainment-seeking customers in the *information marketplace* [Dertouzos, 1997]. The central questions are whether the proliferation of new communications channels will lead to fragmentation and diversity of media audience, and will this audience take a more active role as a participant and producer in the communication process.

The Web has high potential as a medium for delivering news information, which has traditionally been one of the main products for media companies. News production ties a large part of the media company's most valuable asset – the expertise of knowledgeable journalists. News product on the Web can be significantly different than its broadcast or print counterpart. More depth and breadth can be provided on the Web than in other media. Online news content can be expanded with background information and explanation, and search tools make vast quantities of related information instantly available to information explorers. But as the first enthusiastic wave of Internet usage is fading, customers are looking for media products that provide them with significantly more sophisticated tools. In the domain of news information, these tools are geared towards overcoming the problems of information overload, and helping people to make sense of news flows that are largely detached from individual and community context.

News organizations are looking eagerly for their position regarding new media opportunities and they are willing to change their production mechanisms and news contents to meet new customer demand. Customization is typically depicted as a fully automated process, where the original news source is *delivered as-is* to the automated software agents for packaging and dissemination of personalized information to end-customers. This approach is somewhat misguided, since these two resources do not rule each other out. There are many ways in which the journalistic work, information professionals (librarians), and software agents can *together* provide a customized news services. For example, it is possible that journalists create multiple versions of the same story already assuming the different viewpoints and needs of the readership. Also the process of creating related contextual information to news stories is clearly not going to be fully automated. Journalists can design multiple versions of contextualization, which are shown to particular users or user groups. Also the conceptual models necessary for good quality content metadata

will need to be created and updated by media professionals. The key to successful services is in combining the skills of professional journalists and the computational media.

Two main forms of customized news services for individuals are

- news focusing services that reduce time and effort needed to know about the news events that relate to the topics of personal interest, and
- news augmentation tools that help individuals to move from low-salience mode of learning of mass media communications to news interfaces that promote individual knowledge creation by customizing the news to the customer's individual and community context.

Personalized focusing services are useful when they save time, but they may sacrifice the diversity and serendipity of information. Therefore, they fit best in covering the needs in subject areas of professional interests, hobbies, and news on geographical areas.

Flexible access to personally meaningful content, like customized news augmentation, enhances understanding of the topics in the news. Traditional newspaper articles favor a narrative structure that flows logically from one item to the other. Stories often build to a conclusion and end with some form of a reference to expected developments in the future, which provides a sense of closure and explanation. In customized news augmentations this narrative structure is combined with a variety of forms of representation, including customized diagrams, graphs, lists, and fact boxes, which offer the reader a large number of entry points to related material. As a result, the role of sensemaker shifts from the news producer to the news customer.

Communities benefit from the customized news services, because

- community-oriented focusing and augmenting of news content is possible,
- · media companies can create content in cooperation with active communities, and
- news services are tailored to help community members to increase the level of dialogue and social learning through discussions that relate to events that are valuable to communities.

News industry has not traditionally catered the special needs of non-geographic communities of interest and communities of practice. These used to be the area of small-circulation newsletters and special interest publications and magazines. Community-based customization provides new ways for content producers to serve communities. A community profile permits tailoring news to all members of the community.

News is an important external source for social and organizational knowledge, and often serves as a basis for dialogue in a social setting. The meaning of the news story is influenced by the immediate spatio-temporal individual context, and social community context. Effective use of distributed collaboration environments requires shared mental models that guide users in sensemaking and categorization.

People are becoming participative members of creative community instead of merely users. They are increasingly communicating with others and creating meaningful things rather than merely "viewing" and watching. Tools that can support information and knowledge creation, access, summarization, sharing, and

visualization are becoming increasingly important. These communication tools make use of customer models that support interaction tailored to the knowledge, interests, and preferences of individuals, and community models that facilitate collaboration and effective communication amongst the community members. One of the biggest challenges facing systems in support of social creativity is to allow end-users to become active producers and co-developers of news content.

Customized news systems and interfaces enable better tools for individual sensemaking and community communications. The opportunity for media companies lies in taking into account the differences between individual consumers, and understanding the possibility to actively fill the media gap by supporting active groups of people, especially communities of practice, that embrace the Web as their communication medium of choice.

List of Figures

Figure 1. The media gap, adapted from [Neuman, 1991]	2
Figure 2. The role of focusing and augmenting in news services	4
Figure 3. The front page of My Yahoo!	10
Figure 4. Selecting the contents for the front page of My Yahoo!	11
Figure 5. Selecting news by topic areas for personal front page	12
Figure 6. Customizing the layout of the front page of My Yahoo!	12
Figure 7. Augmented news story from February 17, 1999 by CNN	15
Figure 8. "Conflict in Kosovo" theme page	16
Figure 9. An interactive map of Kosovo and surrounding areas, and a timeline of Kosovo history	16
Figure 10. Levels of context in news creation and reception	18
Figure 11. Three dimensions of news reporting: depth, breadth, distance	23
Figure 12. Different levels of context in news creation and reception	28
Figure 13. Triangular relationship between the event, communicator, and receiver [Hartley, 1982]	29
Figure 14. Different dimensions of domain ontology [see Part II: Publication 5]	40
Figure 15. Typical lifecycle of a community of practice [Wenger, 1998b]	45
Figure 16. Dimensions of customized news content	49
Figure 17. Anatagonomy/SS (picture from Sakagami et al. [1998])	54
Figure 18. The Remembrance Agent suggests relevant documents based on current document context	58
Figure 19. Autonomy ActiveKnowledge TM recommends related pointers	58
Figure 20. PLUM map augmentation [Elo, 1996]	60
Figure 21. Three dimensions of customer modeling, adapted from [Rich, 1983]	62
Figure 22. Levels of expertise in a hierarchical conceptual model	64
Figure 23. Modular structure of a user modeling system	67
Figure 24. Levels of interest and expertise in a hierarchical conceptual model	69
Figure 25. Modeling the importance of different communities for an individual.	74
Figure 26. The relationships between user profiles, community profiles, and stereotype profiles	75
Figure 27. The news story with relations to structured domain ontology	77
Figure 28. Customer model organized according to metadata dimensions	77
Figure 29. Layered model of augmentation	81
Figure 30. Augmentation system architecture	83
Figure 31. Financial news story with customized augmentations	84
Figure 32. Broadcasting model with local customization	87
Figure 33. Personalized news content is delivered according to an individual profile	88
Figure 34. The infomediator collects news and routes them to communities and individuals	89
Figure 35. The news producer provides content for all members of the communities	90

Figure 36. The community mediator routes news to individuals	90
Figure 37. Community-centric model of customization	91

List of Tables

Table 1. News values [Hartley, 1982]	22
Table 2. Media skills and knowledge related to news [Potter, 1998]	35
Table 3. The analysis of possible community profile contents	73

References

ACHIEVE (1998) The Economics of New Media - Best Practices in Online Publishing, ACHIEVE Project, European Commission Report, 30 May 1998. At http://www.achieve.ch/ACHIEVE_WG2_FinalReport.pdf

Ackerman M. S., Halverson C. (1998) *Considering an Organization's Memory*, ACM Conference on Computer-Supported Cooperative Work (CSCW'98), November, 1998.

Agre, P. E. (1998) "Designing Genres for New Media: Social, Economic, and Political Contexts," in Jones, S.G. (ed.) *Cybersociety 2.0: Revisiting Computer-Mediated Communication and Community*, Thousand Oaks: Sage Publications

Altheide, D.L. (1976) Creating reality: how TV news distorts events, Sage, Beverly Hills, USA

Balabanovic, M. (1998) *Learning to Surf: Multiagent Systems for Adaptive Web Page Recommendation*, Ph.D. Dissertation, Stanford University.

Baldonado, M.Q.W. (1997) An Interactive, Structure-Mediated Approach to Exploring Information in a Heterogeneous, Distributed Environment, Ph.D. Dissertation, Stanford University. At http://www.parc.xerox.com/istl/members/baldonad/

Baldonado, M.Q.W., Winograd, T. (1997) SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests, Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '97), Atlanta, Georgia, April 1997

Bartlett, F. C. (1932) Remembering, Cambridge University Press, Cambridge, UK.

Bateson, G. (1979) Mind and Nature: A Necessary Unity, Bantam Books, New York.

Belkin, N. J., Oddy R. N., Brooks, H. M. (1982a), "ASK for information retrieval: Part I. Background and theory," *The Journal of Documentation*, Vol. 38, No. 2, June 1982.

Belkin, N. J., Oddy R. N., Brooks, H. M. (1982b), "ASK for information retrieval: Part II. Results of a design study," *The Journal of Documentation*, Vol. 38, No. 3, September 1982.

Belkin, N., Croft, B. (1992) "Information Filtering and Information Retrieval: Two Sides of the Same Coin," *Communications of the ACM*, Vol. 35, No. 12.

Bender W. (1996) "Community and personalization", IBM Systems Journal, Vol. 35, No. 3&4.

Bender W., Chesnais P., Elo S., Shaw A., Shaw M. (1996) "Enriching communities: Harbingers of news in the future", *IBM Systems Journal*, Vol. 35, No. 3&4

Berghel, H. (1999a) "Value-Added Publishing," Communications of the ACM, January 1999, Vol. 42, No. 1

Berghel, H., Berleant, D., Foy, T., McGuire, M. (1999b) "Cyberbrowsing: Information Customization on the Web," *JASIS: Journal of the American Society for Information Science*, Vol. 50, No. 6

Booch, G., Jacobson, I., Rumbaugh, J. (1997) The Unified Modeling Language User Guide, Addison-Wesley, New York.

Bourdieu, P. (1998) On Television, The New Press, New York.

Boyle, C., Encarnacion, A. O. (1994) "An Adaptive Hypertext Reading System," User Modeling and User-Adapted Interaction 4(1).

Brewer, W. F., Nakamura, G. V. (1984) "The nature and functions of schemas," in R. S. Wyer Jr. and T. K. Srull (eds.), *Handbook of Social Cognition*, Vol. 1, Erlbaum, Hillsdale, NJ.

Brown, J. S., Duguid, P. (1991) "Organizational Learning and Communities-of-Practice: Toward a Unified View of Working, Learning, and Innovation," *Organization Science*, 2(1).

Brown, J. S., Duguid, P. (1992) "Enacting Design for the Workplace," in Adler, P. S. and Winograd T. (eds.), *Usability: Turning Technologies into Tools*, Oxford University Press, Oxford, UK.

Brown, J. S., Duguid, P. (1996), "Keeping it simple," in Winograd, T. (ed.), *Bringing Design to Software*, Addison-Wesley, Reading, MA.

Brusilovsky, P. (1996) "Methods and techniques of adaptive hypermedia," User Modeling and User Adapted Interaction, Vol. 6, No. 23.

Carey, J.W. (1986) "Why and How? - The Dark Continent of American Journalism," in *Reading the News*, R. K. Manoff and M. Schudson (eds.), Pantheon Books, New York.

Chen, P. P. (1976) "The entity-relationship model – Towards a unified view of data," ACM Transactions on Database Systems, Vol. 1., No. 1, January 1976.

Chesnais, P., Mucklo M., Sheena J. (1995) *The Fishwrap Personalized News System*, Proceedings of the 2nd International Workshop on Community Networking, Princeton, NJ.

Clancey, W. J. (1997) Situated Cognition, Cambridge University Press, Cambridge, UK.

Darnton, R. (1975) Writing News and Telling Stories, Daedalus, Spring.

De Bra, P. (1999) *Design Issues in Adaptive Web-Site Development*, Proceedings of the 2nd Workshop on Adaptive Systems and User Modeling on the WWW. At http://www.contrib.andrew.cmu.edu/plb/WWWUM99_workshop/

De Kerckhove D. (1997) Connected Intelligence: The Arrival of the Web Society, Somerville House Publishing, Toronto, Canada

Denning, P. (1982) "Electronic Junk," Communications of the ACM, Vol 25., No. 3.

Dertouzos, M. (1997) What Will Be: How the New World of Information Will Change Our Lives, HarperCollins Publishers, New York

Devlin, K., Rosenberg D. (1996) Language at Work, Analyzing Communication Breakdown in the Workplace to Inform System Design, CSLI Publications, Stanford University, United States

Dimitrova, N. (1999) "Digital Video for the Time Impaired," IEEE Multimedia, April-June, 1999.

Donath, J. (1995) Visual Who: Animating the affinities and activities of an electronic community, ACM Multimedia 95, San Francisco, CA

Driscoll, J. (1996) *What is News?*, News in the Future consortium meeting, MIT Media Laboratory, May 1996.

Driscoll, J., Turpeinen, M., Bender, W. (1997) *Silver Stringers and their Melrose Mirror*, draft, available at <u>http://nif.media.mit.edu/papers/silver.pdf</u>

Elo, S. K. (1995) Augmenting Text: Good News on Disasters, DAGS'95: Electronic Publishing and the Information Superhighway, Boston

Elo, S. K. (1996) *PLUM: Contextualizing News For Communities Through Augmentation*, Master's Thesis, MIT Media Lab, Cambridge, MA

Engelbart, D. (1962) Augmenting Human Intellect: A Conceptual Framework, Summary Report, Stanford Research Institute, Stanford

Engelbart, D. C. (1995) *Toward Augmenting the Human Intellect and Boosting our Collective IQ*, Communications of the ACM, Vol. 38, No. 8

Fischer, G. (1998) *Beyond 'Couch Potatoes': From Consumers to Designers*, Proceedings of the 3rd Asia Pacific Computer Human Interaction Conference, IEEE Computer Society, 1998, pp. 2-9

Fuller, J. (1996) News Values, The University of Chicago Press, Chicago

Goldfarb, C. F., Prescod, P. (1998) The XML Handbook, Prentice Hall, Upper Saddle River, NJ

Graber, D. (1989) Processing the News, Longman, New York

Greer, J., McCalla G. (1993) *Student modeling: the key to individualized knowledgebased instruction*, NATO ASI Series F, Vol. 125, Springer-Verlag, Berlin

Gruber, T. (1993) "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," in Guarino, N., Poli, R. (eds.) *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Kluwer Academics Publishers, Netherlands

Gunter B. (1987) *Poor reception, misunderstanding and forgetting broadcast news*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA

Hagel J., Rayport J.F. (1997) *The coming battle for customer information*, Harvard Business Review, January-February 1997, pp. 53-65

Hartley, J. (1982) Understanding News, Methuen, London, UK.

Heinonen, A. (1999) Journalism in the Age of the Net. Changing Society, Changing Profession. Ph.D. Thesis, Tampere University, June 1999

Huang X., McCalla G.I., Greer J.E., Neufield E. (1991) "Student model," User Modeling and User-Adapted Interaction, Vol. 1, No. 1, Kluwer Academic Publishers, Netherlands

Höijer, B. (1992) "Reception of television narration as a socio-cognitive process: a schema-theoretical outline," *Poetics 21*.

Höök, K., Karlgren, J., Wærn, A., Dahlbäck, N., Jansson, C. G., Karlgren, K., Lemaire B. (1996) A glass box approach to adaptive hypermedia, User Models and User Adapted Interaction 6

Höök, K., Wærn, A., Rudström, Å (1997) *Edited Adaptive Hypermedia: Combining Human and Machine Intelligence to Achieve Filtered Information*, Flexible Hypertext Workshop, The Eighth ACM International Hypertext Conference (Hypertext'97)

Jennings, N., Wooldridge, M. (1998) Agent Technology: Foundations, Applications, and Markets, Springer-Verlag

Jensen, K. B. (1986) Making sense of the news, Aarhus University Press, Aarhus, Denmark.

Jones, S.G. (1998) "Information, Internet, and Community: Notes Toward an Understanding of Community in the Information Age," in Jones, S.G. (ed.) *Cybersociety 2.0: Revisiting Computer-Mediated Communication and Community*, Thousand Oaks: Sage Publications

Katz, J. (1994) "Online or Not, Newspapers Suck," Wired, 2.09, September 1994

Kay, J., Kummerfeld, R. (1996) User Model Based Filtering and Customisation of Web Pages, Fifth International Conference on User Modeling, Hawaii.

Keane, P., Mooradian, M. (1998) *Digital News: Delivery Strategies to Meet Changing Consumer Habits*, Jupiter Communications, October 1998

Lassila, O., Swick, R. (1999) *Resource Description Framework (RDF) Model and Syntax Specification,* W3C Architecture Domain, <u>http://www.w3.org/TR/REC-rdf-syntax</u>

Lave, J. (1991) "Situated Learning in Communities of Practice," In L. Resnick, J. Levine, & S. Teasley (eds.), *Perspectives on Socially Shared Cognition, American Psychological Association*, Washington, DC, pp. 63-82.

Levitt, T. (1980) "Marketing Success Through Differentiation of Anything," *Harvard Business Review*, January/February, 1980, Vol. 58, Issue 1, Reprinted in Levitt, T. (1986) *The Marketing Imagination*, Free Press, pp. 72-93

Levy, M. R. (1992) *Learning from Television News*, in Cook, P.S., Gomery D. Lichty L.W. (eds.): *The Future of News*, 1992, Woodrow Wilson Center Press, Washington, D.C.

Lieberman, H. (1995) *Letizia: An Agent That Assists Web Browsing*, Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, August 1995.
Maes, P. (1994) "Agents that Reduce Work and Information Overload," *Communications of the ACM*, Vol. 37, No. 7., July 1994.

Malone T. W., Grant K.R., Turbak F.A., Brobst S.A, Cohen M.D. (1987) "Intelligent information-sharing systems," *Communications of the ACM*, Vol. 30, No. 5.

Marriott, M. (1998) "Happy Birthday, Your Team Won, Your Stock Crashed", *New York Times*, September 17, 1998

Mathé, N., Chen J. (1996) Usercentered indexing for adaptive information access, User Models and User Adapted Interaction 6.

McQuail, D. (1993) Mass Communication Theory, 3rd edition, Sage, London

Milosavljevic, M. (1997) Augmenting the User's Knowledge via Comparison, in Jameson A., Paris C., and Tasso C. (eds.), User Modeling: Proceedings of the Sixth International Conference, UM97, Vienna, New York; Springer Wien New York

Mings, S. M. (1998) Uses and Gratifications of Online Newspapers: An Audience-Centered Study, Ph.D., Rensselaer Polytechnic Institute.

Minsky, M. (1975) "A framework for representing knowledge," in Winston, P. (ed.), *The Psychology of Computer Vision*, McGraw-Hill, New York.

Mostafa, J., Mukhopadhyay S., Lam, W., Palakal, M. (1997) A Multilevel Approach to Intelligent Information Filtering: Model, System, and Evaluation, ACM Transactins on Information Systems, Vol 15, No. 4, October 1997, pp. 368-399

Moukas A., Maes P. (1998) "Amalthaea: An Evolving Multiagent Information Filtering and Discovery System for the WWW," *Journal of Autonomous Agents and Multi-Agent Systems*, Vol. 1, No. 1.

Neuman, W. R. (1991) The Future of Mass Audience, Cambridge University Press, UK.

Neuman, W. R. (1986) The Paradox of Mass Politics, Harvard University Press, Cambridge, MA

Nonaka, I., Takeuchi, H. (1995) The Knowledge-Creating Company, Oxford University Press, New York.

Orr, J.E. (1990) "Sharing Knowledge, Celebrating Identity: Community Memory in the Service Culture", in *Collective Remembering*, Middleton D. and Edwards, D. (eds.), Sage, Newbury Park, CA.

Orwant, J. (1996) For want of a bit the user was lost: Cheap user modeling, IBM Systems Journal, Vol. 35, No. 3&4.

Pazzani, M. (1999) "A Framework for Collaborative, Content-Based and Demographic Filtering," AI *Review*, to be published.

Pilar da Silva, D., Van Durm, R., Duval, E., Olivié, H. (1998) *Concepts and documents for adaptive educational hypermedia: a model and a prototype*, Proceedings of the 2nd workshop on Adaptive Hypertext and Hypermedia, Pittsburgh, USA, June 20-24, 1998.

Pine II, B. J. (1992) *Mass Customization : The New Frontier in Business Competition*, Harvard Business School Press, Cambridge, MA.

Postman, N. (1985) Amusing Ourselves to Death – Public Discourse in the Age of Show Business, Penguin Books, New York

Potter, W. J. (1998) Media Literacy, Sage Publications, Thousand Oaks, CA

Reeves, B., Nass, C. (1996) The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places, CSLI Publications, Cambridge University Press, Cambridge, UK

Rennison, E. (1994) Galaxy of News: An Approach to Visualizing and Understanding Expansive News Landscapes, Proceedings of UIST 1994, Marina Del Ray, CA.

Rennison, E. (1995) Personalized Galaxies of Information, Proceedings of CHI'95, Denver, CO

Resnick, P., Varian, H.R. (1997) "Recommender Systems," *Communications of the ACM*, Vol. 40., No. 3, March 1997

Rheingold, H. (1993) *The Virtual Community: Homesteading on the Electronic Frontier*, HarperCollins Publishers, New York.

Rich, E. (1983) "Users are individuals: Individualizing user models," International Journal on Man-Machine Studies 18.

Russell, D. M., Stefik, M. J., Pirolli P., Card, S. K. (1993) *The cost structure of sensemaking*, Proceedings of the ACM SIGCHI Conference on Human Factors in Computing (INTERCHI '93), Amsterdam, Netherlands, April 1993.

Saari, T. (1998) Knowledge Creation as the Production of Individual Autonomy. How News Influences Subjective Reality, Licentiate thesis. University of Tampere, Finland

Sack, W. (1994) On the Computation of Point of View, Proceedings of the National Conference of Artificial Intelligence (AAAI 94), July 31-August 4, 1994, Seattle, WA.

Sakagami, H., Kamba, T. (1997) *Learning Personal Preferences On Online Newspaper Articles From User Behaviors,* Sixth International World Wide Web Conference Proceedings.

Sakagami, H., Kamba, T., Sugiura, A., Koseki, Y. (1998) *Effective personalization of push-type systems* — *visualizing information freshness*, Proceedings of the 7th World-Wide Web Conference, Brisbane, Australia

Savia, E., Kurki, T., Jokela S. (1998) *Metadata Based Matching of Documents and User Profiles*, Proceedings of the 8th Finnish Artificial Intelligence Conference.

Schank, R. C., Abelson R. P. (1977) Scripts, Plans, Goals and Understanding, Lawrence Erlbaum, New Jersey, USA

Schank R., Abelson R.P. (1995) "Knowledge and memory: The real story," *Advances in Social Cognition*, Volume VIII, Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA

Seybold, P. (1998) Customers.com, Random House, New York.

Shardanand, U., Maes, P. (1995) Social Information Filtering: Algorithms for Automating "Word of Mouth", Proceedings of CHI'95 Conference on Human Factors in Computing Systems, ACM Press

Shoemaker, P. (1991). *Gatekeeping; Communication concepts series (Vol. 3)*, Sage Publications, Newbury Park, CA.

Stadnyk, I., Kass, R. (1992) "Modeling User's Interests in Information Filters," *Communications of the ACM*, Vol. 35, No. 12, December 1992.

Suchman, L. (1987) Plans and Situated Actions, Cambridge University Press, Cambridge, UK

Södergård, C., Aaltonen, M., Hagman, S., Hiirsalmi, M., Järvinen, T., Kaasinen, E., Kinnunen, T., Kolari, J., Kunnas, J. (1999) *Integrated multimedia publishing - combining TV and newspaper content on personal channels*, 8th International World Wide Web Conference, Toronto.

Tapper, J. (1995) "The ecology of cultivation: a conceptual model for cultivation research," *Communication Theory*, Vol. 5, No. 1, February 1995.

Taylor, A. G. (1992) *Introduction to Cataloging and Classification*, Libraries Unlimited, Englewood, Colorado:, 1992.

Tomita, T. (1980) *The New Electronic Media and Their Place in The Information Market of the Future*, in Anthony Smith, ed., Newspapers and Democracy: International Essays on a Changing Medium, pp. 49-62, MIT Press, Cambridge, USA

Weitzman, L., Wittenburg, K. (1994) Automatic Presentation of Multimedia Documents using Relational Grammars. In Proceedings of Second ACM International Conference on Multimedia. 15-20 October, San Francisco. 443-451.

Wellman, B. (1999) "Living networked in a wired world," *IEEE Intelligent Systems*, January/Febryary 1999

Wenger, E. (1998a) Communities of Practice: Learning, Meaning, and Identity, Cambridge University Press, Cambridge, UK

Wenger, E. (1998b) "Communities of Practice: Learning as a Social System," *Systems Thinker*, Vol. 9, No. 5 June/July 1998

Wexelblat, A. (1999) *History-Based Tools for Navigation*, Proceedings of the Hawaii International Conference On System Sciences, January 5-8, 1999, IEEE Press

Wiley, D. L. (1998) "Beyond Information Retrieval: Ways to Provide Content in Context," *Database*, Vol. 21, No. 4, August/September 1998.

Winograd, T, Flores F. (1986) Understanding Computers and Cognition: A New Foundation for Design, Ablex, Norwood, NJ.

Wittgenstein, L. (1953) Philosophical Investigations, Blackwell, Oxford, UK

Woodruff, E., Chakravorty, M., Smith Lea, B. (1996) "The Role of Technology in Educational Reform," In K. Leithwood (Ed.) *The Knowledge Base for Educational Reform: Recommended Research Priorities.* Final Report to the Ontario Ministry of Education and Training. Toronto, Ontario.

World Wide Web Consortium (1996) Cascading Style Sheets, level 1. At <u>http://www.w3.org/TR/REC-CSS1/</u>

World Wide Web Consortium (1997) *Platform for Internet Content Selection (PICS)*. At <u>http://www.w3.org/PICS/</u>

World Wide Web Consortium (1998) *Extensible Markup Language (XML)*, W3C Architecture Domain. At <u>http://www.w3.org/XML/</u>

World Wide Web Consortium (1999a) *Web Content Accessibility Guidelines 1.0.* At <u>http://www.w3.org/TR/WAI-WEBCONTENT/</u>

World Wide Web Consortium (1999b) *Platform for Privacy Preferences (P3P) Specification*, W3C Working Draft. At <u>http://www.w3.org/TR/WD-P3P/</u>

Part II

Summary of Publications

The following summary briefly introduces the publications in Part II of this thesis and the author's contribution to each publication.

Publication 1.

Turpeinen, M., Saarela, J., Korkea-aho, M., Puskala, T., Sulonen, R. (1996). *Architecture for Agentmediated Personalised News Services*, PAAM'96: The First International Conference on Practical Application of Intelligent Agents and Multi-Agent Technology, 22-24 April 1996, London, UK.

Publication 1 presents a system architecture for news filtering. This architecture was used in building a personalized version of the news material available in the Ota Online project at Helsinki University of Technology. The introduced system combined metadata-based matching with keyword-based matching and social filtering. The author was the main author of the paper and presented the paper at the conference.

Publication 2.

Saarela, J., Turpeinen, M., Korkea-aho, M., Puskala, T., Sulonen, R. (1997). *Logical Structure of a Hypermedia Newspaper*, Information Processing and Management. Vol. 33, No. 5, pp. 599-614.

Publication 2 introduces publication templates, structured articles, and procedural style sheets for multipurpose publishing. The use of SGML format called News Industry Text Format (NITF) and style sheets was experimented to produce multiple versions of the content. The author was the principal co-author of the paper that covers many of the core concepts of the Ota Online project.

Publication 3.

Kurki, T., Jokela S., Turpeinen M., Sulonen R. (1999). Agents in Delivering Personalized Content Based on Semantic Metadata, Intelligent Agents in Cyberspace, Papers from the AAAI Spring Symposium, Technical Report SS-99-03, AAAI Press.

Publication 3 discusses the SmartPush system for filtering the information flow to provide a personalized news service to news customers. Personalization and delivery process was modeled as adaptive agents, and the metadata structures were used for matching the news content as well as learning about user interests. The author participated in the SmartPush research team in 1998-99, and in this paper contributed mostly to the areas of metadata creation, agent architecture, and discussion on the agenthood of SmartPush agents.

Publication 4.

Turpeinen, M. (1999). *Augmenting Financial News for Individuals and Organizations*, Special Issue on Intelligent Agents for Education and Training Systems, International Journal of Continuous Engineering Education and Life-Long Learning.

Publication 4 depicts a prototype of a Financial News Augmentation system that combines continuous information streams with selected data from heterogeneous information sources. This paper also emphasizes the role of metadata in the process and introduces the combination of individual and community models for information augmentation. The author was the sole writer of the paper.

Publication 5.

Jokela, S., Turpeinen, M., Sulonen, R. (2000). *Ontology Development for Flexible Content*, HICSS-33 Minitrack on Systems Support for Electronic Business on the Internet, January 4-7, 2000.

Publication 5 discusses the challenges of ontology development and describes a model for combining the ontology development to the electronic publishing process. The paper is mainly a result of discussions and experience in the SmartPush, and with experiments at Alma Media Corp. The author was the principal co-

author of the paper, and presented the paper at the conference. The paper received the best paper award of the *Internet and Digital Economy* track of the HICSS-33 conference.