The Role of Structured Content in a Personalized News Service

Sami Jokela
Helsinki University of Technology / Andersen Consulting, Center For Strategic Technology Research (CSTaR)
sami.jokela@hut.fi

Marko Turpeinen
Alma Media Corp.
marko.turpeinen@almamedia.fi

Teppo Kurki
Helsinki University of Technology / Alma Media Corp.
teppo.kurki@kauppalehti.fi

Eerika Savia
Helsinki University of Technology / Done Wireless Oy
eerika.savia@done360.com

Reijo Sulonen
Helsinki University of Technology
reijo.sulonen@hut.fi

Abstract

Digitalization of content and exponential growth of Internet and electronic commerce are changing the media industry. The availability of structured content enables new ways to produce and deliver information. This paper explains the role of semantic metadata in developing content for an adaptive news service in the SmartPush -project. In SmartPush, news content is categorized using semi-automatic tools and pre-defined vocabularies. Metadata enhanced content is then matched against user profiles to provide customers with a personalized news service. After providing the personalized news to the customer, SmartPush system adapts the personalization based on user feedback.

This paper discusses the requirements of personalized content services and challenges in an approach based on structured metadata. We describe how supporting ontologies for the content were developed and maintained and what kinds of tools were developed to support the structured metadata creation. We also present some results of the pilot phase of the project and introduce some of the issues observed during the system implementation and in the performed field trial.

1 Introduction

News has traditionally been offered as a ready-made package that journalists produce according to the production rhythm of their media. The proliferation of digital media content is fundamentally changing the practice of news production. The production pace is also increasing as more and more news organizations are producing news 24-hours a day, seven days a week, and new types of news medium have emerged in the marketplace. Most notably the Web has rapidly developed into a major news medium in the 1990’s with the following distinguishing characteristics:

- multimedia content creation: multiple media journalism, requiring new talents from news producers,
- content customization: tailoring content 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 and usage are possible.

Without carefully designed news services and advanced content applications, the customer will likely suffer from the effects of information overload and information decontextualization. To help alleviate these problems, news content can be tailored dynamically to individuals and communities. In principle, there are three potential approaches to customize a news service:

- content can be selected, grouped, tailored, and organized according to customer preferences and interaction,
- 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.

Personalized news content services are useful when they save time and reduce the amount of irrelevant information, but they may also sacrifice the diversity and serendipity of information. Therefore, they fit best in covering the long-term needs in well-defined subject areas such as professional interests, hobbies, and news on geographical areas.

For the purposes of customized content delivery, semantically rich content descriptions can be used to gather dynamically changing user models as well as to match the news content against these customer profiles. Semantic metadata, defined as information about the meaning of the content, can be created automatically or semi-automatically. Whichever the case, quality assurance of metadata and ontology creation requires human professionals.

The introduction of metadata in the publishing process will lead to large organizational changes at the news producers, if they are to provide attractive customized services. Customer profiles and high quality metadata become increasingly valuable when news organizations are facing new competition. New intermediaries, such as Web portals, are also collecting the necessary information for customization to meet the special needs of individuals and customer communities.

This work has concentrated on structured and human-edited metadata, instead of fully automated content metadata such as inverted indices or term vectors used in text information retrieval. Structured metadata for customized news services is motivated by research that states that domain models are central to understanding, and especially hierarchical categorizations are important in making the connections between the meaning of the incoming news stories and the concepts understood by an individual[14]. Although these conceptual models are fluid and negotiated in social interaction, they form a basis for making sense of the news events. Therefore, there is a need for an explicit domain model, which is used when deep multi-dimensional metadata is created for customized news services.

Several areas of research, such as digital libraries, information retrieval and integration, knowledge management, as well as artificial intelligence, have examined methods and tools for describing knowledge in a formal model (see e.g. [1, 3, 11]). To a certain extent, the role of the news industry is to create and maintain these models for the news content domain, covering the topics that are classified as news. News media thus creates a “map of news landscape” for their customers. One role of customized media is to understand how this map is modified to best meet each customer’s needs.

This paper describes the SmartPush project, where the key to success was to creatively combine the skills of professional journalists with automated software tools for metadata creation and content customization. The main goal of SmartPush was to manage information overflow by focusing the news customer’s attention to a subset of news content by filtering and prioritizing the content. Filtering emphasizes the function of leaving out unnecessary pieces of information from constantly available and evolving information streams (see e.g. [3]). In SmartPush an automated system filters news based on adaptive customer profiles. The profiles store customer interests and adapt to customer feedback on the news. Prioritization depicts information in a manner that highlights the most relevant information to the individual in a personalized way.

Key challenges in the implementation of a personalized news service are (1) describing the content as metadata, (2) modeling the changing short-term and long-term information needs of the users, and (3) providing suitable architecture for information brokering.

Customers have different short-term and long-term interests, which are modeled in the system. Tracking individual’s interests regarding news 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 clearly specify their interests regarding news.
- Casual users are not willing to spend much effort in explicitly specifying their interests.
- Representative sample data covering all the user’s actual interest areas is hard to get.
- The domain ontologies that are used to describe personal interests andthe available content change over time.

Explicit profile manipulation is the simplest method to produce a user interest profile. 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 requires the user to go through and learn additional steps of functionality typically at the early stages of usage when the user is not at all familiar with the system.

Dynamic user profiles capture user interests via implicit or explicit feedback. With implicit feedback the system collects user interests indirectly by monitoring the interaction between the user and the application. Learning algorithms make the profile adapt over time more closely to the viewer's habits. Mostafa et al. [9] show that a news filtering system that detects shifts in interests significantly improves filtering results. Learning system that monitors user's actions should also consider temporal dependency of user interests.

1.1 Related work

fishWrap [2] was one of the first prototypes for personalized newspapers using profiles of individual members of MIT community. fishWrap selects news from interest areas 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, 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. It has filtering services for both individuals and corporate community customers. NewsEDGE relies on semiautomatic metadata creation for vast quantities of news items. Semiautomatic creation means that although most of metadata is machine-generated, human editorial staff assures the quality of metadata.

In most information filtering systems agents analyze the information in complete and original form and create an index for document matching [9, 10]. SmartPush uses an architecture, where a matching agent compares metadata descriptions of incoming material against user profiles as new material is published. The computational requirements for the matching process are therefore greatly reduced.

Content-based filtering is also used in push-oriented services, such as PointCast Network. 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 based on the channels they are subscribed to.

Currently there is a multitude of news producers, such as My CNN and Wall Street Journal Interactive Edition, and intermediaries, such as My Yahoo!, that provide news filtering for individuals at different levels of sophistication. 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 [8]. However, the main shortcomings of commercially available personalized news services, such as My Yahoo! or Pointcast, are in the following areas:

- Adaptivity. The service has typically no support for dynamic learning or interest profile adaptation 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. The news service uses a simple pre-determined taxonomy of content areas. The categories are also fully defined by the service 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 and products are assumed to remain unchanged over a long period of time. There is no clear mechanism for versioning the categories. Additionally, production systems and methods do not often allow flexible product modifications resulting from changes in customer needs or available media platforms. In fact, the service provider gets easily stuck with the outdated categories and products it started with.
- Multiple usage contexts. There is no possibility for multiple user roles. The customer profile stays the same no matter what the current context of the user is. Customer's needs might be very different if he or 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 often lack necessary 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, because the reader is seeing news based only on such information he or she has been interested in earlier. 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.
- 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.

All of these issues have come up also during the work in the SmartPush project. Some of them, such as adaptivity, were incorporated already in the initial design, whereas others, like product management, came only later from the cooperation with the participating media companies. Some of these issues were not directly visible in the pilot implementation, such as serendipity and multiple usage contexts, but nonetheless they were taken into account in the theoretical work behind the implementation.

2 SmartPush project

SmartPush was a three-year long research project conducted at the Helsinki University of Technology TAI Research Center. The project ended at the spring 2000 and was performed in a close co-operation with a number of industrial partners including Alma Media, Fujitsu TeamWARE, ICL, Nokia, Sanoma-WSOY, and Sonera. In the project content originating from media companies was categorized using pre-defined ontologies to enable advanced processing of the content during its delivery. The metadata represenation of the content was then matched against user profiles to provide the users with a personalized news service on multiple media platforms (Figure 1).

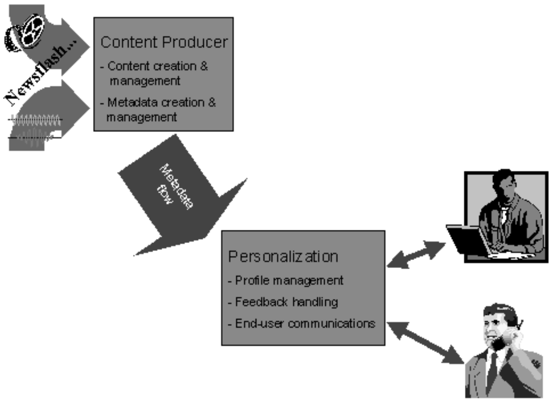


Figure 1. SmartPush overview

The SmartPush project was highly dependent on the availability of metadata-enriched content. Although metadata can describe different qualities of the content such as its format or production related information [5], SmartPush project focused specifically on the semantic metadata. Semantics in the context of SmartPush meant machine-usable descriptions of the important qualities of the content that can express both the characteristics of the content and the interest of users and that are interpreted similarly by both the creator and the user. The project was based on the belief that by structuring and defining the important aspects of content and by limiting the descriptions to a certain domain and detail level, it is possible to define semantic structures that describe sufficiently the content in question. With the help of semantic metadata the SmartPush system was able to track and respond to user's interests in the form of personalized news services and adjust user profiles based on user feedback.

When SmartPush started in 1997 we assumed that suitable standards and tools for describing content were available and in use. However, further research in the field showed that such standards did not exist and there were no tools in wide use that supported our view of describing content with semantic metadata. The lack of suitable standards and methods resulted in putting a considerable effort in creating both tools and models for metadata.

3 Domain ontologies in SmartPush

Before metadata can be produced, we need a metadata model, i.e. a *domain ontology*, which formalizes and structures the content domain. In the SmartPush context, domain ontology means a set of formally specified conceptual structures modeling the semantics of the content.

Ontology comprises a set of concepts and concept relationships representative to the content domain

(Figure 2). Concepts and their relations define conceptual models for classifying information objects under different *dimensions*. Another closely related term to metadata dimension popular in information retrieval is a *facet*. Taylor [15] defines facets as “clearly defined, mutually exclusive, and collectively exhaustive aspects, properties, or characteristics of a class or specific subject.”

By augmenting the document with semantic metadata organized in a collection of dimensions, the content provider describes the qualities of the content from multiple viewpoints. The content provider uses typically a set of conceptual dimensions that have been defined and standardized, such as MARC for library resources, and Dublin Core for network resources [16]. These dimensions include not only subject, but also other dimensions such as author, publisher, and publication date. With these dimensions the ontology should cover such semantics of the content that are needed to produce and deliver content to the customer.

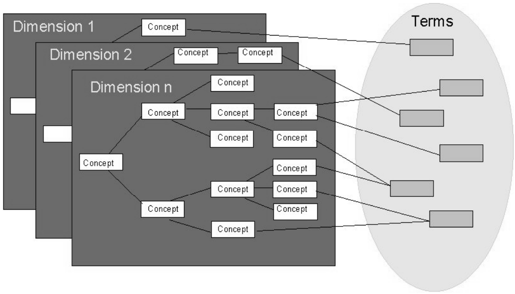


Figure 2. Different dimensions of domain ontology

Metadata dimensions in SmartPush, such as *subject matter* and *media type* [12], were considered to be independent from each other. However, there are often subtle interrelationships between different dimensions that can be difficult to model. For example, *geography* dimension might consist 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).

In news customization, for example with adaptive news filtering, interrelated dimensions significantly complicate the problem of automatically learning and adjusting user’s interests and expertise. If a person is planning to buy 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 car manufacturers, the Japanese news events might have more importance.

The internal structure of a dimension can be flat, hierarchical, graph, or consist of some other complex structure. In a flat structure the concepts are not interrelated. An example of a flat dimension could be the author of a book or keywords assigned to an article. Hierarchies are a common way to organize and manage information. They offer an intuitive way to organize, summarize and navigate large amounts of information. A representative hierarchical dimension could be geographical location, where the world could be divided for example into continents, and then further into countries, states, and cities.

Visualization of the metadata structures sets additional challenges. There are moderately good visualization methods for tree structures up to medium size. If the size of the structure is large or the relations between the concepts in the structure are complex, other kind of structures such as graphs and visualization methods such as hyperbolic trees [7] might be needed.

3.1 Initial metadata efforts in SmartPush

In the initial phase of SmartPush we analyzed the suitability of existing metadata standards for describing content semantics in SmartPush. Although some alternatives were found, those standards did not meet our requirements due to their limited applicability for our purposes, poorly defined semantics, incompatible structure, or lack of support with the content providers. We continued this work by analyzing some proprietary solutions and noticed that some companies had been able to standardize their own de facto formats within their content supply chain. The lack of open standards together with the fact that our project partners had done some work in defining metadata structures in house motivated us to try to define a simple metadata structure ourselves. However, the latest development in the metadata standards field indicates that even the companies with successful proprietary standards are participating in international standardization and are aiming at interoperability with standards such as PRISM and NewsML [17].

The first domain model was built in an early phase of the project. To be able to test the first prototype a substantial amount of structured content was needed. The initial content set consisted of Finnish newswire articles from 1995 and 1996 covering a variety of news categories such as domestic and international financial news and short reports on accidents. We created the metadata model by reading all the articles and collecting central concepts used in the content, after which the metadata model was structured and the actual metadata was created for the selected news articles.

We used the test data set and its metadata in the first two prototypes. The initial data and domain model was suitable for experimenting the various software tools, but it lacked realism and the model was not applicable to domain specific news feeds. The data set was not sufficient for testing the functionality of the personalization since we had no data on the user behavior and thus, no real profiles could be constructed. To evaluate the SmartPush concept in real life we needed a pilot environment that provided us enough realistic structured content and user data.

4 Kauppalehti -pilot

After comparing alternative sources for content the project decided to use real time news material from the online version of the Finnish financial newspaper Kauppalehti. There were multiple reasons for selecting this publication as a test material for SmartPush. Kauppalehti Online, the online department of the paper, produces between 100 and 250 online news items every day, it has a large user base, and it produces content for multiple products and media platforms such as headlines on mobile phones or newsflashes to be incorporated as part of other content products. We also had the possibility to access the online service's pseudonymized log data and, most importantly, the source provided the necessary resources for creating the metadata.

4.1 Domain model for the pilot

Earlier co-operation with Kauppalehti had produced an analysis of the online news publishing process and an initial semantic metadata model for financial news. We used these results as a basis for the domain model and refined it further jointly with content experts at Kauppalehti. The group set initially some guidelines for the dimensions and concepts to be included, after which the content experts in the company drafted the first version of the model and we gave feedback about it. When feedback was incorporated into the model, a content matter expert from Kauppalehti Online began using the model by producing semantic metadata for the incoming news content. The final version of the domain ontology was refined together with the content matter expert based on the experiences gathered from the testing with the actual news material. Details on the principles and methods applicable to building ontologies for content can be found in our previous work [6].

Five distinct dimensions of metadata were defined for SmartPush: *Subject, Location, Industry, Company* and *Priorit*y. Two of them had hierarchical structure, *Subject* describing the content of the document and *Location* that expresses which geographical areas are involved. The content description model, *Subjec*t, was a generalization hierarchy that grouped related concepts together.

SmartPush relied in its personalization on the idea that for each dimension every article contained the same amount of relevance that was divided among the different concepts in the article. Relevance distribution meant that if the article contained multiple concepts belonging to a certain dimension, the overall interest was divided between these aspects and thus we were not able to state that a certain article had more relevant information than others. To tackle this challenge the concept of article *Priority* was introduced. Its intended role was to be an overall category that the reporters were able to use to prioritize the article. However this information was not available on our material.

At the beginning of the pilot a number of changes were done to mostly to the *Subject* dimension, e.g. a number of new nodes were added and some sub-hierarchies shifted under another upper level concept. We did also some structural work on the *Location* dimension. When the actual categorization started, the only modifications to the ontology were the additions of new companies to the *Company* dimension.

Personalization in SmartPush used weights to express the relative importance of different concepts. Weights gave us gained more expressive power compared to a simple binary belongs-to relation. The weights were cumulated upwards in the hierarchy so that the weight of each node was the sum of the weights of its child nodes. This way, both the summarizing categories of the topmost level and the details of the leaf nodes were available for personalization.

4.2 Metadata set

Metadata was produced for about 3700 news articles that were published during a period of three months, which was a subset of the total article feed over that time. Some articles were ignored as data about users reading them could not be acquired. Some recurring summary news items were also ignored. The effort was also concentrated to get metadata for the articles that test users had read or discarded during their online sessions. Semantic metadata related to each article had quite a lot of variance, as each document had a unique set of concepts and dimensions. Only 28% of the documents had metadata in all four dimensions, *Subject*, *Location*, *Industry*, and *Company*. 84% of the articles had metadata in at least three dimensions and 99.8% in at least two dimensions. Table 1 shows the percentage of documents having metadata in each dimension and Table 2 contains the weight distribution of the different dimensions.

Table 1. Comparison of metadata in different dimensions

Dimension	Articles with metadata
Subject	*99.9 %*
Location	*78.6 %*
Industry	*86.0 %*
Company	*46.4 %*

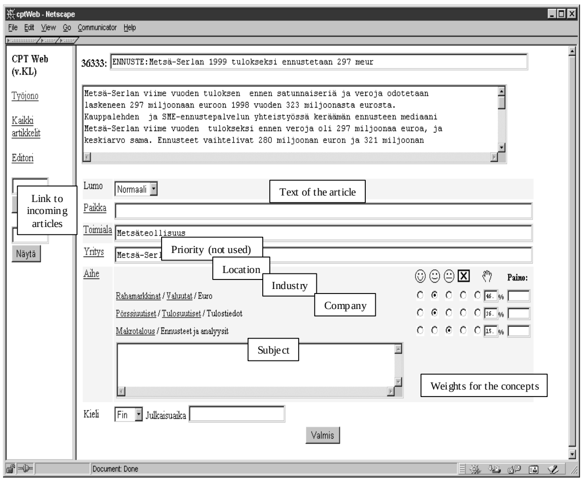
Table 2. Metadata distribution in different dimensions

Dimension	Total number of concepts	Average number of weights
Subject leaf	*95*	*2.5*
Subject top	*8*	*1.5*
Location leaf	*215*	*5.5*
Location top	*7*	*1.3*
Industry	*16*	*1.2*
Company	*227*	*2.1*

4.3 Tool support for metadata production

Many of the current content management systems treat content and metadata creation separately and position the metadata creation only later in the workflow. Although in some cases certain characteristics of the content and publishing process, such as the timeliness of content or content syndication, prevent from creating content and detailed metadata descriptions within the same process step, these two activities should be performed in tight interaction. The original author of the content is the best source to state, what the content is about and what qualities are important in it. If the interaction between content authoring and metadata creation is not ensured or if we rely on automation to take care of the metadata creation, the resulting metadata may suffer from inferior quality. Similarly, without close interaction between the annotator and the content source as well as without experience, clearly stated instructions and tool support, the inter-annotator agreement on appropriate metadata can be low and the quality of metadata may vary greatly between annotators [4].

However, as timeliness is an important quality of content and as human work is extremely expensive, there is clearly room for automation of the routine tasks involved. Fully automatic classification and information extraction from free text is often an unrealistic goal, so the project decided to emphasize intelligence augmentation instead of using fully automated classification tools and built a keyword based metadata authoring tool for the authors. We built and tested different types of content tools and levels of automation during the project and ended up with a web-based implementation illustrated in Figure 3.



36333: ENNUSTE:Metsä-Serlan 1999 tulokseksi ennustetaan 297 meur

Metsä-Serlan viime vuoden tuloksen ennen satunnaiseriä ja veroja odotetaan laskeneen 297 miljoonaan euroon 1998 vuoden 323 miljoonasta eurosta. Kauppalehden ja SME-ennustepalvelun yhteistyössä keräämän ennusteen mediaani Metsä-Serlan viime vuoden tulokseksi ennen veroja oli 297 miljoonaa euroa, ja keskiarvo sama. Ennusteet vaihtelivat 280 miljoonan euron ja 321 miljoonan

Lumo: Normaali

Paikka:

Toimiala: Metsäteollisuus

Yritys: Metsä-Ser

Aihe:

Rahamarkkinat / Valuutat / Euro	○	●	○	○	○	46. %	
Pörssiuutiset / Tulosuutiset / Tulostiedot	○	●	○	○	○	36. %	
Makrotalous / Ennusteet ja analyysit	○	○	●	○	○	15. %	

Kieli: Fin Julkaisuaika:

Figure 3. Metadata production with the Content Provider Tool

The tool, called Content Provider's Tool (CPT), allowed the reporter to add semantic metadata to the news articles.

With CPT a reporter initiates metadata creation by typing a news article into a normal web page or retrieving it from an external source. After that CPT processes the textual body of the article and generates a metadata suggestion that the reporter verifies and modifies by removing and changing the metadata entries through the CPT user interface.

Most of the news articles in the Kauppalehti testing were in Finnish, which unfortunately has a complex morphology rendering the traditional word stemming algorithms in most cases obsolete. The project therefore opted for a more advanced solution in creating the metadata suggestion for the news content and used a morphological analyzer TWOL from Lingsoft [1] to extract nouns from the news stories as terms. Each concept has a group of terms mapped to it and similarly each term might have more than one concept mapped to it. Each mapping has a weight expressing how strong the binding between the term and concept is. When all the mappings for the content are analyzed, we have a list of concepts with weights that are based on the term frequencies and the weights of the corresponding mappings. The highest ranked concepts and their weights are then selected to be the metadata candidates for the document.

The term-concept mappings were generated manually with an administration tool. The manual tool was acceptable for our testing purposes, but it should be enhanced with automation if the administration tool is used for production purposes.

The goal of the metadata authoring tool was to have a simple and extensible prototype of a system that helps in generating metadata. Simplicity was in line with the

[1] www.lingsoft.fi

findings of Belkin and Croft, who stated that simple word-based representations combined with appropriate retrieval models are surprisingly effective as well as being efficient and straightforward to implement [1].

Our tool was able to make metadata suggestions reasonably well, but due to its simplicity it made occasionally naive mistakes. These problems lead to modifications in the keyword mappings of our domain model. For example, keywords causing confusion were removed and others missing from the initial mappings were added. Some kind of changes were expected, however, as whole the idea with mappings was to allow flexibility to the metadata structures while they are being used.

The metadata creation process was later improved by adding more automated tasks based on simple pattern matching to take care of routine tasks. Very short stories and recurring summary type stories were automatically ignored, and a number of automated routine metadata entries such as name identification simplified and speeded up the task of entering metadata.

4.4 Content flow in SmartPush

Content feed from Kauppalehti was received via email in XML News format [18], which has a simple structure for describing news items. The received metadata was initially relatively scarce, but it was improved during the piloting to contain company and industry metadata. Although our metadata augmentation was not part of the original content authoring process, our metadata editor worked in tight and constant interaction with the content sources. With the CPT tool the metadata editor was able to pick up an article, receive a suggestion for the content metadata including the metadata originating from Kauppalehti, make the necessary corrections and complete the augmentation by sending the metadata further to the personalization and delivery. This information flow in the pilot can be seen in Figure 4.

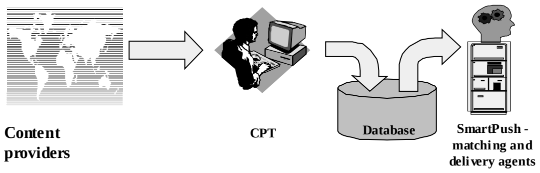


Figure 4. Metadata augmentation with the Content Provider Tool

Once the metadata was produced it was sent to the matching agent in the SmartPush system to make it available to the users as fast as possible. All the metadata was also stored for later usage. The matching agent compared the metadata with existing user profiles and submitted the results over to the delivery agent, which managed the delivery of results to the users. These results were used to provide the users with ranked lists of all the news items published during a single day.

We needed an end-user interface to be able to test the functionality of the personalization. The overall design goal for the user interface was not to alienate the potential users so we ended up having an HTML–based interface similar to the interface Kauppalehti Online users normally use for reading online news. The most notable difference in our interface was that it used frames to keep the topic list visible when user read some news item. This way we could provide means for giving feedback to the article after it was read.

The recommendations were presented as a simple sorted list with the same layout as in the original Kauppalehti Online interface. The news list could be sorted in five different ways: one priority-ordered list for each metadata dimension and one list ordered by the time the news article had been published. All the pilot users had profile information that was initiated with their earlier reading habits.

4.5 Initial results of the trial

When we analyzed the results we found clear differences in the importance of different dimensions. Some users' interests are better described with relevant industries or companies, whereas others' interests can be more easily expressed with certain geographical locations or subject of the document.

The requirement to keep dimensions rigid and independent from each other was a controversial goal. On one hand rigidity and independence make the computing task easier, but on the other hand we might have to limit the expressiveness of the ontology. Rigidity means in practice that we have to define all supported dependencies already in the ontology structure and we cannot change the structure on the fly. If dimensions were allowed to have flexible structures or dependencies to other dimensions, these qualities would quite likely lead to extremely extensive computations and problems with managing and understanding the internal dependencies within the ontology.

An important requirement for a metadata dimension is that it discriminates the documents so that it is possible to sort them in decreasing order of interest according to a user profile. If multiple distances between documents are equal, the model is not able to discriminate the documents well enough. The lack of discriminate power caused a problem with some of the distance measures we tested during the trial.

The results from the trial with Kauppalehti Online were encouraging although more in-depth research with our test material and usage logs is necessary to draw any ultimate conclusions. We are currently working on this, and these results will be available separately. Another way to gain confidence in the methods would be to have information covering a longer period of time. User logs and metadata for a period of at least one year would help to really see how well the personalization works in practice.

5 Summary and conclusions

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 journalists, information professionals, and software agents can together provide a customized news services. The conceptual models necessary for good quality content metadata will need to be created and updated by media professionals.

When SmartPush project started in 1997, we excluded domain modeling and semantic metadata production from our research plan and assumed that standards and methods for semantic metadata are readily available. However, we had to revise our plan quite soon after realizing that our assumptions were false. We experimented with different metadata models and production tools and finalized a prototype system in the fall of 1999 that was used to produce enhanced content with semantic metadata.

The idea of using semantic metadata for information filtering is a relatively novel idea. The research in personalization lead to propose hierarchical structures for semantic metadata as well as asymmetric methods for comparing the profile information and the documents [13]. With a hierarchical metadata structure it is possible to represent information in various degrees of detail. More importantly, hierarchical structures assist to draw top-level summarizations of the metadata and to extend the impact of metadata to the neighboring concepts allowing approximations with a smaller set of available metadata. Hierarchies seemed also to be a suitable method to reduce the amount of required calculations in the personalization as well as an intuitive way to navigate a multitude of concepts during the concept creation as long as the ontology structure in general is not overly complex.

Objective analysis on how well the semantic metadata describes the content is difficult and laborious to conduct. Even subjective analysis would, unfortunately, require a great amount of manual work. Consequently, we did not concentrate on analyzing the descriptiveness of the metadata. Although we have discussed the different components affecting semantic metadata quality in our previous work [6], the development of suitable methods to measure metadata quality should be addressed in the future research.

The semantic metadata for actions was not defined in the SmartPush domain model. By concentrating only on nouns in the mappings we effectively excluded the notion of changes taking place in the news article. Latter work with domain modeling has showed that action is in some cases one of the main reasons why people are interested in the news.

It is difficult to express various levels of representation with a single set of metadata. SmartPush news articles were simple in the sense that they generally did not require multiple levels of representation. However, with more complex content, such as books, the corresponding metadata needs to be defined for different levels of the content expressing what are the metadata descriptions for a paragraph, a chapter, and the whole book. If we just simply accumulate the metadata from the lowest level upwards, we end up with overly complicated and detailed descriptions on the top level and cannot benefit from the advantages of strong prioritization on the top level.

Semantic metadata tools seemed to work well for the metadata production, but by putting more time and extending the level of automation on refining the mappings between keywords and concepts the manual effort could have been reduced. However, this would have been possible only after the ontology itself had been stabilized and the users would have been familiar with the ontology as well as gained confidence on the reliability of the system.

Content provider tool could also have been improved with functionality that learns from the encountered categorization problems. The system could monitor the modifications and corrections the reporter does to the automatically created metadata suggestions and then reflect the changes to the keywords and their mappings.

According to our experiences with the pilot testing metadata-based personalization does work, and structured semantic metadata can be produced roughly as outlined in the SmartPush project. For time critical information the metadata production has to be integrated with the actual content production. Personalization by itself may not justify this level of effort, but if there are other uses for metadata, personalization can be seen as an added benefit.

Semantic metadata can be used together also with other types of content than text. Pictures as well as audio and video clips, which are becoming more and more available in the news feeds, could be enhanced with semantic metadata, although the automatic metadata extraction is more challenging. Once the semantic metadata is available, a variety of new metadata-based services and operations such as personalization, advanced content management, and content reuse on multiple products and media platforms become possible independent of the type of the original content. Similarly, if corporations are willing to open and describe their operating domain, semantic metadata and personalization could be used to link internal knowledge-related processes to external content sources.

Analysis on the possibilities and limitations of the advanced content applications and services could motivate media companies to develop their semantic metadata production and services like SmartPush, and should therefore be included as a potential and interesting topic for future research in the field.

6 Acknowledgements

The authors wish to thank SmartPush project sponsors Finnish Technology Development Centre TEKES and Alma Media, Fujitsu TeamWARE, ICL, Nokia, Sanoma-WSOY, and Sonera as well as their current employers for providing resources to continue the research.

7 References

[1] Belkin, N. J., Croft, W. B. (1992) Information Filtering and Information Retrieval: Two Sides of the Same Coin? Communications of the ACM, December 1992, Vol. 35, No. 12.

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

[3] Foltz, P. W., Dumais, S. T. (1992) Personalized Information Delivery: An Analysis of Information Filtering Methods, Communications of the ACM, December 1992, Vol.35, No.12.

[4] Hirschman, L., Brown, E., Chinchor, N., Douthat, A., Ferro, A., Grishman, R., Robinson, P., Sudheim, B. (1999) Event 99: A Proposed Event Indexing Task for Broadcast News, Proceedings of the DARPA Broadcast News Workshop, Herndon, Virginia, February-March 1999, http://www.itl.nist.gov/iaui/894.01/proc/darpa99/pdf/dir5.pdf

[5] Jokela, S., Saarela, J. (1999) A Reference Model for Flexible Content Development, Proceedings of The Second International Conference on Telecommunications and E-Commerce (ICTEC'99), Nashville, Tennessee, October, 1999.

[6] Jokela, S., Turpeinen, M., Sulonen, R. (2000) Ontology development for flexible content, Proceedings of the Internet and the Digital Economy Track of the 33rd Hawaii International Conference on System Sciences (HICSS-33), Maui, Hawaii, January, 2000.

[7] Lamping, J., Rao, R. (1996) Visualizing Large Trees Using the Hyperbolic Browser, Conference on Human Factors in Computing Systems (CHI 96), Vancouver, Canada, April, 1996, www.acm.org/sigchi/chi96/proceedings/viedos/Lamping/hb-video.html

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

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

[10] 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.

[11] Nonaka, I. (1991) The Knowledge-Creating Company, Harvard Business Review, November-December 1991, reprinted in Harvard Business Review on Knowledge Management, pp. 21-45, Harvard Business School Press, Boston, Massachusetts, 1998.

[12] Savia, E., Kurki, T., Jokela S. (1998) Metadata Based Matching of Documents and User Profiles, Proceedings of Human and Artificial Information Processing, 8th Finnish Conference on Artificial Intelligence (STeP'98), Jyväskylä, Finland, September, 1998.

[13] Savia, E. (1999) Mathematical Methods for a Personalized Information Service, Master's Thesis, Helsinki University of Technology, http://smartpush.cs.hut.fi/pubdocs/

[14] 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.

[15] Taylor, A. G. (1992) Introduction to Cataloging and Classification, Libraries Unlimited, Englewood, Colorado, 1992.

[16] Weibel, S., Miller, E. (2000) WWW-site for Dublin Core Metadata Element Set, http://purl.org/metadata/dublin_core

[17] XML Europe (2000) Extensible Markup Language (XML) conference. Paris, France, June, 2000, http://www.gca.org/papers/xmleurope2000/

[18] XMLNews (2000) WWW-site for the XMLNews specification, http://www.xmlnews.org/