Ontology Structures in CoMet

Jukka Yli-Koivisto and Juha Puustjärvi

Software Business and Engineering Institute, Helsinki University of Technology, P.O.Box 9600, FIN-02015 Helsinki University of Technology, Finland
{Jukka.Yli-Koivisto, Juha.Puustjarvi}@cs.hut.fi

Abstract. The distribution of personalized electronic publications is based on the compatibility of user profiles and document's metadata. Ideally an ontology is a common denominator for user profiles and document's metadata. Hierarchical ontologies are flexible in that document's metadata may have different granularity compared to user profile. This means that they may be defined on different levels of the hierarchy. This in turn makes the matching of document metadata and user profiles more complex. In this paper we describe the CoMet System, and in particular how the matching problem is solved.

1 Introduction

Various publishing houses have started to exploit new publishing medias such as Internet and PDA-devices. In these environments publishing houses encounter new challenges. More content can be distributed compared to the conventional newspapers where space is limiting the actual information quantity available. In electronic publication explanatory background material can be linked to news items. This is called *information augmentation* and it provides larger perspectives to readers. Several information sources must be utilized in providing background material from various heterogeneous document management systems. Document *metadata* describes the content of the documents, which are stored in document management systems. Information augmentation is therefore dependent of the services provided by the system that stores the metadata.

Selective dissemination of information (SDI) is a form of electronic publishing. Newspapers that are available online are good examples of SDI services. Online newspapers are already very common. According to WebWombat [12] there are 66 online newspapers at the moment in Finland.

User profile is knowledge about a user and his or her interests. User profiles are compared with document's metadata. If a match is found the document will be delivered to the customer. Therefore content management independent metadata handling model must be implemented to provide the needed services for the news publishing industry.

Our solution for the problem of handling various heterogeneous document management systems in SDI is to build a content management independent metadata handling system with meaningful domain ontology. This is achieved by separating metadata from its original content. The separate metadata database called

metadatabase is handled by using the facilities provided by the conventional relational databases. The developed system will benefit from traditional database services such as data abstraction, high-level access through query languages and controlled multi-user control. The system can also be distributed to geographically different locations and yet remain efficient. The core system serves multiple push and pull type applications at the same time through its service interface.

User and document profiles are stored using hierarchical domain ontology. The granularity of profiles can be different. This increases the complexity of document delivery decision i.e. matching. Our solution removes this problem by introducing the LCH-matching method that is refined to weighted LCH-matching. Document similarity is calculated using *largest common hierarchy* (LCH). The finding of LCH can be done efficiently by using database technology.

The remainder of the paper is organized as follows. In Section 2, we look at the related work in electronic publishing and particularly in online newspapers. Then, in Section 3, we introduce the basic problem area and the environment of the *CoMet System*. In Section 4, we consider the hierarchical ontology structure of the CoMet System, and in particular the way the matching problem is solved. Finally, in Section 5, we present the conclusions and the need for further research.

2 Related Work

SmartPush [10, 14] is a personated delivery system for economic news items and it has similarities to our work. SmartPush used a hierarchical ontology that is similar to our profile construction. However the used similarity calculation method differs from our approach. SmartPush uses *asymmetric similarity measure* [15] that doesn't exclude documents from the result set. It rather ranks the incoming document flow to have a certain order in a result set. This is clearly a disadvantage according to our aims in information augmentation and *data re-use*. SmartPush uses agents in its implementation. Our approach differs also in this section. We will exploit relational database and its calculation power for matching purposes. SmartPush used a matching agent that was coded in Java. Our vision is that a relational database can offer a rather efficient way of handling huge document metadata mass compared to calculation power that Java can offer.

Fishwrap [5] is a personalized newspaper that uses news material from several external news providers. It allows topic selection and layout customization of the personalized news page. Incoming news feed is matched against the user specified topics. With this pre-categorizing the system gains in performance compared to online matching of all topics. Fishwrap has also community wide features. Page one contains several news items. Fishwrap keeps count on the interest of news items by counting the number of readers that a news item has. News item's position on the page one changes according to the popularity it has gained. Fishwrap has also information augmentation features. Fishwrap checks its photo and sound databases for pictures and sound recordings that mach to the news item.

User behavior had a significant role in user profile adaptation in Krakatoa Chronicle [3] and especially in its successor project Anatagonomy [9]. The used Java

applet enabled very intensive customer behavior surveillance. The customer behavior is tracked while she or he reads: activities like scrolling, maximizing, opening articles in new windows or saving them into scrapbook is assumed to reflect positive interest towards the article content. The storing method for implicit user profiles wasn't described in the article but the used calculation methods and user profile characterizing indicates similar keyword list presentation like in explicit user profiles. Krakatoa Chronicle looks like an ordinary newspaper because it has a multi-column layout and justified text. News items relevance was shown as a slider widget that the customer could re-adjust for feedback. Krakatoa Chronicle and Anatagonomy used client-server architecture where a server handed the news items to the client, which was responsible for layout generation and feedback surveillance. This kind of architecture doesn't suit to the CoMet System because it limits the used customer terminals.

Telepublishing [7] was implemented in HyperNeWS. This method enables a newspaper like layout as in Krakatoa Chronicle but the used implementation method restricts all the terminals that do not have HyperNeWS available. A second layout type was also available. It was designed to support the noted ways of customer's electronic newspaper usage. A significant amount of work was targeted to develop ways to support the content creation process with electronic newspapers. One notable feature in Telepublishing is background material that is offered to news items. This feature has similarities to our information augmentation feature in the CoMet System. However Telepublishing seems to offer its background material in a limited amount compared to our System. Personating features are also present but the implementation of matching problem is not clarified in the article that describes the Telepublishing system.

3 CoMet - environment

CoMet stands for ***co***ntent management of a media company based on ***met***adata. We continue the work that has been started in our predecessor project called SmartPush. The aim of the project is to build a working prototype of the CoMet System and test it in real content production environment. The System will provide various services that help content producers daily activities. The CoMet System can be used as the main intelligence of personalized push and pull services that will be distributed to different terminals including mobile phones and desktop computers. Our aim is to design metadata and ontology structures and their distribution that can be used in electronic publishing.

Content re-use and information augmentation are key issues in electronic publishing [13]. The benefits of content re-use are obvious. If content has already been produced it would be waste of recourses to produce it again. Content authoring is efficient only if content is produced only once. In information augmentation explanatory background material is added to content. This can be seen in electronic newspapers as links to relevant background material. In this way the content of an electronic publication can be enriched to meet the customers personalized needs.

3.1 CoMet Solution Issues

We must admit that a publishing house cannot abandon their current content production environments and simply start using a new one. Current systems could have gone through intensive customization to meet the desired production and business needs. Therefore a separate guide to information must be presented. Metadata must be stored in a relational database along with the location of the document. The CoMet System provides a transparent access to all documents within a production company. All media types like text, video and sound can be distributed by our system.

Our focus is to provide services for electronic publishing and especially for the news publishing industry. Our System will be a relational database centric solution for information handling. The CoMet System gains from the benefits of relational databases such as data abstraction, high-level access through query languages and controlled multi-user control. Publishing houses can have geographically distributed content management systems. The CoMet System can be distributed as well to meet the needed performance demands and for hiding the actual content production point.

In SDI systems information is distributed according to user profiles that contain the information about users interest areas. Therefore a method for matching documents and user profiles must be introduced [6]. The vector space model is a popular way to calculate the similarity of a user profile and a document. It has been alleged that recall and precision of the *vector space model* is superior compared to the *Boolean* (relational) *model* [4].

We suggest that the vector space model is not the only suitable matching method in the electronic publishing environment. This is due the limited size of source collection involved in electronic publications. Especially in news distribution a news item is relevant only for a short period of time. We argue that an efficient SDI model can be specified by SQL-queries with a relatively high precision and recall. To be able to do this user profiles and document metadata must be presented in such fashion that effective matching can be calculated by the relational database management system.

3.2 CoMet System

Electronic newspaper is a combination of news items, documents, pictures and other media objects produced by a publishing house and external news providers. Externally produced content must be converted and stored to publishers internal document management systems. The produced content can then be distributed to end-users. Different terminals such as desktop computers and mobile phones can be used for reading the content that has been made available. The CoMet System handles the distribution and personating of content. Overview of the CoMet System architecture can be seen in figure 1.

The CoMet System handles distribution and personating duties in a publishing house. *CoMet Kernel* acts as an information mediator between information sources and their users. It provides services for content creators, editors and other customers using SDI services. Content distribution is based on created metadata. This

information is compared to user profiles. If a document is found to be interesting for a user according to his or her user profile, the document is show to him or her with relevant background material.

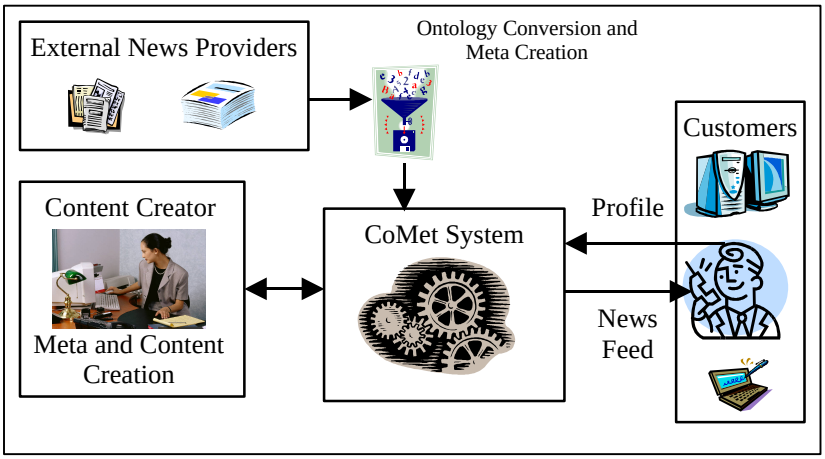

Fig. 1. CoMet architecture

Document metadata usage is the key element in the CoMet System. Therefore it must be aware of all the documents that are saved to the content management systems within a publishing house. Documents can still be placed to the various document management systems but a gateway to them must be implemented. With document management system gateways the CoMet System offers transparent access to all documents. The CoMet System utilizes these gateways when documents are delivered to end-users.

3.3 CoMet Kernel

The CoMet Kernel is responsible for all the personalizing and information augmentation features provided by the CoMet System. The main purpose of the CoMet Kernel is to provide services to applications within a publishing house that need personalized SDI services for information distribution. *CoMet Intelligence* that constructs the core of CoMet Kernel does this all. Figure 2 shows the details of the CoMet System. CoMet Kernel is the center part of the CoMet System. It contains CoMet Intelligence, document metadata database (Meta DBMS in figure 2) and document store.

The metadata database has an essential role in CoMet Kernel. It contains document metadata and user profiles. Documents are stored to a separate document management system i.e. document store. This is due the fact that publishing houses are dependent on their current solutions for document management. This is because of the business functions that are built heavily on the document systems that are currently being used

in content production. The used document management systems already provide good versioning functions and authoring facilities. Therefore documents are separated from metadata. Location of a document is stored in metadata. CoMet Kernel provides a gateway to current document management systems. Therefore current document creation processes can remain unaltered. The document gateway can be implemented to file systems (FS), object databases and other sources.

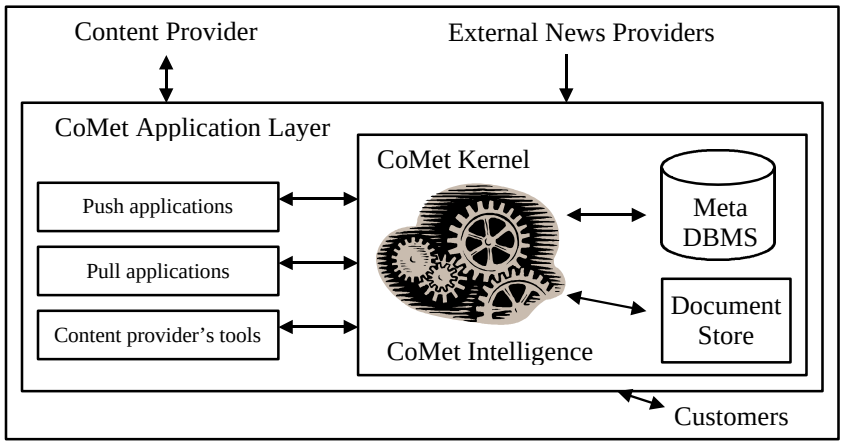

Fig. 2. The CoMet System

3.4 CoMet Application Layer

CoMet System divides into two main parts: CoMet Kernel and *CoMet Application Layer*. The latter is the part of the system that interacts with the users, like content creators and customers. Applications have a service interface that they can use. Through this interface applications in CoMet Application Layer can interact with the CoMet Kernel. Applications have the basic service portfolio in use and they do not have to worry about SDI issues. The main concern is the amount of information augmentation used and the wanted layout. CoMet Kernel does not provide layout issues at all. The layout is done in the application layer. Therefore content is separated from the presentation to be able to support different terminals [13].

CoMet Kernel can simultaneously serve several CoMet Application Layer applications. Basically this interface works as follows. CoMet Kernel provides a line of services. These services can be called to provide SDI functions. Service request can be for example a document request for a customer who wants all the news items relevant to his or her profile. CoMet Kernel does the matching and then delivers the documents to requesting application.

3.5 Provided Services

The CoMet System has four main services. Firstly, we look at the personalizing features. The CoMet System has metadata information about customers and documents. The CoMet System is capable of performing information filtering task and acting as an information mediator between information sources and their users. Secondly, the CoMet System has content re-use services. If a news item that has been originally created for a television newsreader is used again in electronic publication, re-using of content is utilized. The CoMet System can assist content creators in finding relevant related news items and other documents. Thirdly, the CoMet System provides information augmentation features. In information augmentation explanatory background material is added to content. This can been seen as links to documents that create relevant background material for a news item. Fourthly, *story chain management* can be used. This information about articles and their relations against each other is stored in metadata database.

4 Metadata Structures and Matching

Metadata has very important role in the CoMet System. The main purpose of metadata is to contain a compact representation of a document's content and a user profile. Once metadata is created from the actual document content, it should be possible to process it independently from the original content. This knowledge is used later on to deliver relevant documents to the CoMet System customers without accessing the original documents. Created metadata information must be machine-readable. This means that metadata must be processed without human assistance. Metadata creation doesn't have to be fully automated and it can contain manual phases [14].

Used metadata structures must be implemented in such fashion that a uniform metadata format can be used to describe content from different sources. A structured data format, that can contain different value types, is an efficient way of implementing metadata structures. We use XML [16] for presenting the CoMet metadata structures. This kind of structured data is easy to handle and different accessing methods like DOM (Document Object Model) [17] has been introduced for handling it. However our approach leaves the chosen metadata standard open as long as a structured metadata format is used.

Metadata usage has several advantages over using document content for deciding if a document is distributed to a CoMet System's customer [8]. The obvious advantage is the size of metadata. It captures the essential semantics of the source. Therefore created metadata is potentially smaller than the actual document. Size intensive file types like video and sound benefits form the use of metadata because it can save a significant amount of storage space and computation time. Metadata supports all media formats by using only one representation format. This is a significant advantage because only one matching algorithm must be implemented to cover all media types and their distribution. The only disadvantage is the time that must be consumed for metadata creation. Another difficult issue is the need for changing

metadata structures. New media formats may require changes in the used metadata structures. In this case it is not clear how old and new metadata formats relate to each other.

4.1 Hierarchical Ontology

Ontology can be understood in many ways depending on the circumstances it appears. By the ontology we mean a set of metadata structures consisting of concepts and their relations against each other. These concepts build a definition of the problem domain. The ontology can contain different angles, i.e. *dimensions* of the content. These dimensions contain the needed information for describing the documents within the problem domain. Each dimension describes one aspect of the problem domain such as story subject, author or geographic location of the story. Every concept in a dimension is orthogonal, i.e. it is independent from other concepts in other dimensions. Dimensions build up from concepts and their relation. They form a hierarchical model of a dimension.

Ontology and their structures can vary depending on implementation. We will use *hierarchical ontology* structures because of the calculation and the implementation advantages. Besides that hierarchical model is efficient to handle and it provides easily interpreted visual description of dimensions and their concept relations. Human observer can quickly form an impression of the used dimension and the concept relations that bears within Subject dimension. A simple example of a hierarchical ontology can been seen in figure 3. In this example a hierarchical ontology forms a

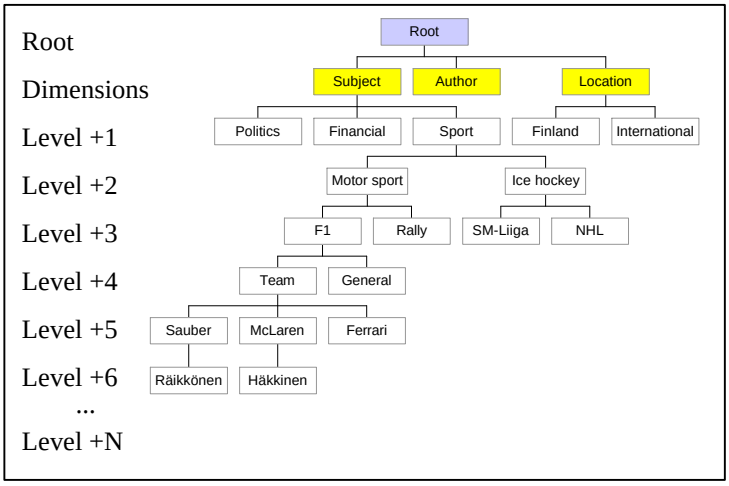

Fig. 3. Ontology construction

tree. The root level is the top node for this hierarchical model. Under that are dimension levels. From that level different orthogonal dimensions can be accessed. Here we have three dimensions: Subject, Author and Location. We take Location into

closer observation. In the next level (level + 1) Location divides into two parts: Finland and International. This is the leaf level of Location dimension. Finland and International are concepts in this dimension.

It is important to notice that ontology structures must be created before metadata. Metadata is in fact the terms that relate to concepts in the dimensions of ontology. Therefore the needed metadata information is presented in hierarchical ontology construction. Ontology creation is a difficult task [8]. A person that has a good knowledge of the problem domain and customer's needs should create the used ontology. Several iteration times in ontology creation could be needed before a useful ontology is found. The problem domain will evolve and change over time. This means that ontology is never in its final state and therefore it must be updated periodically.

4.2 Document Profiles

Document profile contains the semantic metadata that has been created to describe the content of the actual document. Metadata is stored in the hierarchical ontology structures. In figure 4 we can see a news item and its metadata information. This simplified example is based on the Subject dimension that has been introduced in figure 3. The news item is about formula 1 driver Kimi Räikkönen and his thoughts before his first F1 race in Interlagos circuit. Our ontology fits into this problem domain and it is therefore capable of describing this news item. The news item reflects its content to Subject dimensions Sport concept.

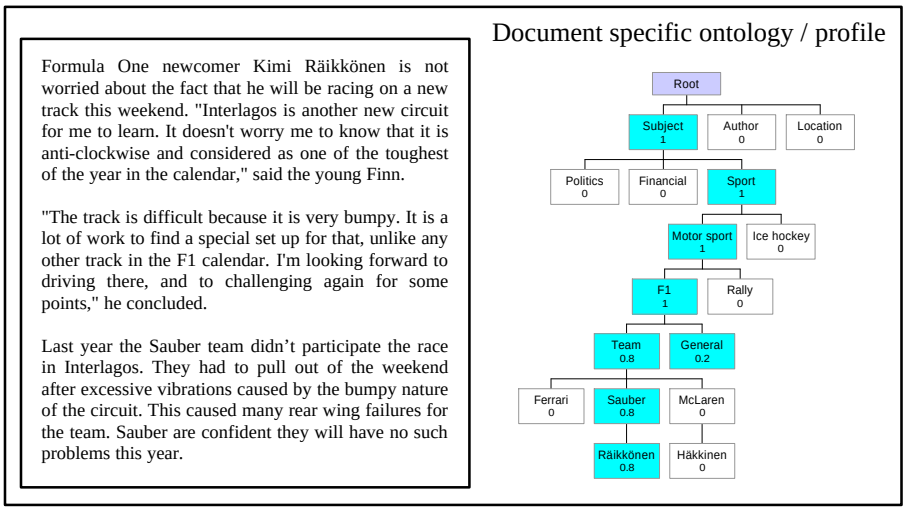

Fig. 4. Document profile

The news items relations to subject dimension concepts is expressed as weights in the leaf nodes of the used hierarchical ontology. Leaf weights are then summarized to its parent node. Parent nodes weight is the sum of its child node weights. This

summarization is continued until dimension level is reached. If weights have been defined to concepts, the dimension level will always have value 1. Every dimension is independent from other dimensions. If the values are not set, the dimension level node will have value 0. This means that every dimension is normalized between 0 and 1. In figure 4 concept F1 has a value 1. This is calculated as a sum from its child node weights 0.8 (Team) and 0.2 (General). The summarization is then continued and finally weight 1 is put to Subject concept in dimension level.

Document metadata creation is a part of document creation process. Content creators or editors are responsible for metadata creation. When we take a look at the news item and its ontology in figure 4, it is clear what the weights in concepts are reflecting. This is because people are very talented in observing document content [11]. Human expertise is therefore needed in metadata creation process. In the SmartPush project a content provider's tool was created to help the metadata creation task [10]. The content provider's tool analyzed texts content by using certain key terms like country names for capturing the needed metadata information. Content creators and editors then modified these generated suggestions for suitable semantic metadata information. The tool helped the basic metadata generation task but the accuracy of the metadata was still in the hands of a human expert.

4.3 User Profiles

A user profile (Figure 5) captures the user interests in machine understandable form. A user profile can be build from a set of keywords that describe the preferred interest areas of a customer. Keywords are compared against news items. If a news item contains a term or several terms from the keyword list, it is shown to the customer. This method is used in Krakatoa Chronicle [3]. Another popular way to store user profile as a term vector associated with term weights [2]. We are using the same hierarchical ontology that was presented earlier. In this way we are able to use the same hierarchical presentation model for documents and user profiles. This can be seen as an advantage when document matching is taken into closer observation.

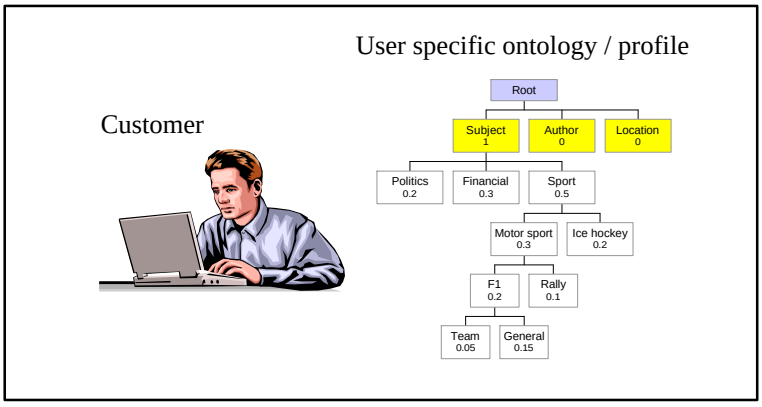

Fig. 5. User profile

In figure 5 we can see a customer and his user profile. It consists of Subject, Author and Location dimensions although only Subject dimension has meaning in this user profile. Subject dimension divides into Politics, Financial and Sports. Sport concept has the finest presentation because that has been refined to level + 4. Concept weights are calculated in the same way as in document ontology. User profile uses the same ontology that was presented in figure 3.

A user profile must be created before a customer can use the SDI services that are provided by the CoMet System. Its purpose is to describe the long-term information needs of a customer in a particular subject area. Customers personal information needs can vary in daily basis but these changes in customer interest areas cannot be captured. Therefore long-term interest is the only interest form that can be captured in a sufficient accuracy. In this method document recall and precision can be guaranteed.

4.4 Matching

The idea of personalized SDI is based on distributing relevant documents to customers depending on their individual information needs. A user profile is utilized when knowledge is needed concerning customer interests. Decision on document delivery is made according the knowledge about document content and its relevance to the user. The matching problem occurs when a decision on document delivery must be made according to customer's user profile. This kind of decision problem exists as well in information retrieval (IR). IR research has raised three retrieval models that are applicable also to SDI. These models are the Boolean model, the vector space model and the probabilistic model [1]. The first two models are used broadly in different SDI implementations.

It is important to notice that Boolean and vector base model don't fit into our environment. This is because our purpose is to distribute and augment information that has various media types like text, video and sound. If SDI is implemented according to these models, the actual document content is analyzed when user profile and document profile matching is calculated. This would be impossible when for example the decision on video material distribution is made. Instead we rely on document metadata and user metadata when matching is calculated.

4.5 LCH-matching

The CoMet System needs a matching method that is capable of handling large amount of metadata material to enable information augmentation and data re-use. We are using metadata information for the matching calculations. This helps us to decrease the calculation burden because metadata information is typically compact compared to document text or video information. Despite of these advantages that we have already gained, the selected matching method still needs to be efficient. Typically SDI research has concentrated on recall and precision problem. These are important aspects but they really don't matter if high recall and precision are achieved with a calculation method that is untenable with its time consumption [18]. Acceptable response times must be guaranteed for information augmentation and data re-use.

We have decided to use the hierarchical ontology structure. Once the ontology has been constructed we can use it for modeling the user profiles and the document profiles. The use of one ontology is an advantage because calculation methods are easier to design and implement when user profiles and document profiles use the same hierarchical model. Time-consuming ontology mappings are not needed and the calculation process is more intuitive for human observer than in the situation where the use and the document profiles use a different ontology.

The CoMet System compares the document profiles against the user profiles and decides the closeness of these two entities. The first step is to find the largest common hierarchy (LCH). It is a definition for the largest hierarchy that the user profile and the document profile share in the used ontology. In this way we can isolate the potentially relevant documents away from the discarded document set. The LCH can be examined as a whole or dimensions can be observed independently. Dimensions can be divided further to concept braches. As we can see from figure 6, a dimension can split into different subjects like Sport and further into Motor sport and Ice hockey. These concept branches form separate subject areas that are relevant for this particular customer. When these areas are inspected separately better augmentation recall and precision can be obtained than in the situation where the whole LCH is treated as one entity. Depending on the service (augmentation, matching data re-use or story chain detection) different matching methods can be used. Therefore the matching granularity can differ from a service model to another. This enables us to fine tune the matching process depending on the task that the CoMet System is working on.

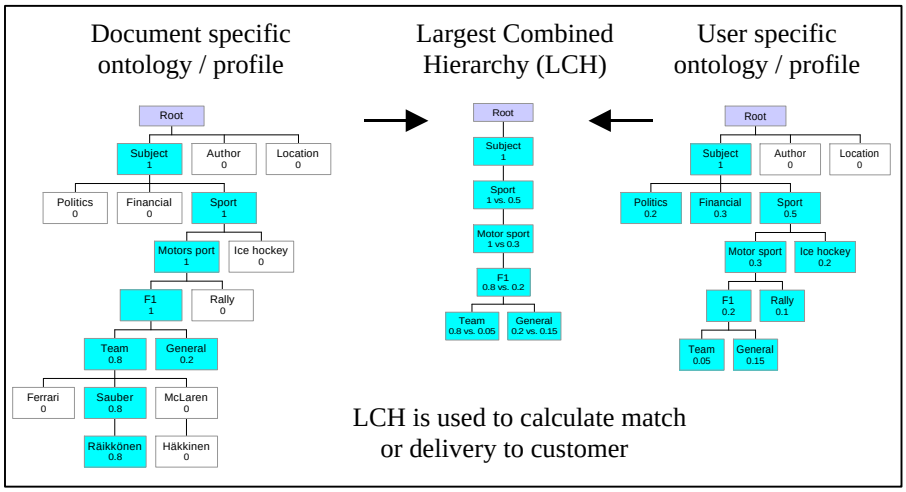

Fig. 6. Largest common hierarchy matching

Next we present a simplified example of the LCH detection and the final matching calculation method. In figure 6 we have two profiles: the user profile and the document profile. When these two profiles are inspected the largest common hierarchy can be found in the center of figure 6. This is rather efficient inspection

because hierarchical trees can be easily examined. The resulted LCH has only one branch but this doesn't have to be the case. In fact in our case similarities can be found only from one dimension but in more complex situations LCH reaches over several dimensions. The final stage involves the actual calculation. We can see that calculation is relevant only from level +1 and deeper in the hierarchy because dimension level is always 1 for both profiles. Calculated elements would be Sports (1 vs. 0,5), Motor sports (1 vs. 0,3), F1 (0,8 vs. 0,2), Team (0.8 vs. 0.05) and General (0.2 vs. 0.15). The result from the comparison of these five concepts bears the final similarity measure in this simplified example. In this manner we have been able to exclude the non-relevant documents from the result set. After that a similarity measure was calculated to those documents that had LCH with a user profile. The result set can be arranged according to the similarity measure results.

4.6 Weighted LCH-matching

The deepness of the used hierarchy has a significant effect on the expression power of the used ontology. This can be seen from figure 3 where Subject dimension is rather deep (level + 6). In level +1 Sport concept is introduced. When we look at the level +3, Sport has been sharpened to F1 concept. If LCH is found from level +3 i.e. F1, it's more significant than LCH that reaches only to level +1 i.e. Sport. The expression power of level +3 is quite strong in our ontology. If we presume that two documents exists which both have LCH with the user profile only in one dimension. The LCH-matching ranks them to be equally similar against the user profile and the LCH is located in the same branch but the other one seems to have deeper level. In these circumstances the document having deeper LCH should be ranked above the other document. This can be arranged by emphasizing those matching results that are calculated in deeper dimension levels.

Besides of excluding documents from the result set, the deepness of LCH can be exploited in inverted fashion. If a user profile is very limited or contains only few deep branches, the CoMet System benefits from user profile generalization. Now we presume that only one branch has been defined for a customer's user profile. This branch is the same as the LCH in figure 6. A customer is obviously very interested in F1 news items. If the first weighted-LCH results exclude all news items from the result set, user profile generalization can be exploited. If we emphasize the similarity calculation weight that is used in deepest level –1, a broader view to the defined user profile ontology can be constructed. If a customer is interested in F1 news items, one can assume that he would like to have other motor sport related news items when F1 news items are not available. This kind of feature could be used as a "see also"-type of recommendation service. Besides showing a relevant document a customer can broaden his or her view by examining the articles that belongs to a nearby concept in the systems hierarchical ontology.

5 Conclusions and Future Work

We have examined the need for selective dissemination of information (SDI) for a publishing company. This problem area contains for example electronic publications like online newspapers. External news providers and internal content creators create news items and other documents such as pictures and sound material. Distribution of a news item is based on metadata information, which is a compact representation of document content.

Then we focused on metadata presentation model. We presented a hierarchically ontology structure for metadata information container. This model is used for storing the customer's user profiles and the document profiles. LCH-matching (Largest Common Hierarchy) is used for matching user profiles and document profiles. News items are distributed according to the calculation result of LCH-matching. Then we specified the LCH-matching method to the weighted LCH-matching. In this matching model the hierarchical ontology is exploited to adjust the matching result. The deepness of the used hierarchy has a significant effect on the describing force of the used ontology. Therefore the deepness of the used LCH must be emphasized in the matching calculations.

Our goal was to define a content management independent metadata handling system that contains the needed information to be able to provide SDI services in electronic publishing. The basic metadata structures and an approximate description of the CoMet System architecture were presented in this paper. The future challenges include a more detailed description of weighted LCH-matching and especially the architectural design of the distributed metadata database. The metadata database will also be responsible for calculating the matching of user profiles and document profiles. This is a very important issue in our future work. According to our research in matching implementations it will be a novel way of doing the matching calculation. Relational database will provide a very powerful facility for handling large amounts of metadata information.

References

1. Belkin, N., Croft W., Information Filtering and Information Retrieval: Two Sides of the Same Coin? Communications of the ACM, Vol. 35, No. 12, December, pages 29 – 38, ACM 1992.
2. Bell, T., Moffat, A., The Design of a High Performance Information Filtering System. In Proceedings of the 19th ACM SIGIR conference on Research and Development in Information Retrieval, August 18 – 22, Zurich Switzerland, ACM 1996.
3. Brahat, K., Kamba, T., Albers, M., Personalized, interactive news on the web. Multimedia Systems, Vol. 6, September, pages 349 – 358, Springer-Verlag 1998.
4. Çetintemel, U., Franklin M., Giles C., Self-Adaptive User Profiles for Large-Scale Data Delivery. In Proceeding of the 16th IEEE conference on data engineering, pages 622 – 633, IEEE 2000.
5. Chesnais, P., Mucklo, M., Sheena, J., The Fishwrap Personalized News System. In Proceedings of the 2nd International Workshop on Community Networking

Integrating Multimedia Services to the Home, June 20 – 22, Princeton, New Jersey USA, pages 275 – 282, IEEE 1995.
6. Foltz, P., Dumais, S., Personalized Information Delivery: An Analysis of Information Filtering Methods. Communications of the ACM, Vol. 35, No. 12, December, pages 51 – 60, ACM 1992.
7. Haake, A., Hüser, C., Reichenberger, K., The individualized electronic newspaper: an example of an active publication. Electronic Publishing, Vol. 7(2), June, pages 89 – 111, Wiley 1994.
8. Jokela, S., Turpeinen, M., Sulonen, R., Ontology Development for Flexible Content. HICSS-33 Minitrack on Systems Support for Electronic Business on the Internet, January 4 – 7, 2000.
9. Kamba, T., Sakagami, H., Koseki, Y., Anatagonomy: a Personalized Newspaper on the World Wide Web. International Journal of Human-Computer Studies, Vol. 46, No. 6, June, pages 789 – 803, Academic Press 1997.
10. Kurki, T., et al., 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 1999.
11. Labrou, Y., Finin, T., Yahoo! As an Ontology – Using Yahoo! Categories to Describe Documents. In Proceedings of the 8th international conference on Information knowledge management, November 2 – 6, Kansas City, MO USA, pages 180 – 187, ACM 1999.
12. Online Newspapers., www.onlinenewspapers.com, 2001
13. Saarela, J., The Role of Metadata in Electronic Publishing. D.Sc. thesis, Acta Polyechnica Scandinavica, Ma 102, 1999.
14. Savia, E., Kurki, T., Jokela, S., Metadata Based Matching of Documents and User Profiles. In Proceedings of the 8th Finnish Artificial Intelligence Conference, pages 61 – 69, Finnish Artificial Intelligence Society 1998.
15. Savia, E., Mathematical Methods for a Personalized Information Service. Master's Thesis, Helsinki University of Technology 1999.
16. W3C., Extensible Markup Language (XML) (Second Edition). W3C Recommendation 6 October 2000, http://www.w3.org/TR/2000/REC-xml-20001006.
17. W3C., Document Object Model (DOM). http://www.w3.org/DOM.
18. Yan, T., Garcia-Molina H., Index Structures for Information Filtering Under the Vector Space Model. In Proceedings of the 10th International Conference on Data Engineering, February 14 – 18, Houston, Texas USA, pages 337 – 347, IEEE 1994.