Metadata Based Matching of Documents and User Profiles

Eerika Savia, Teppo Kurki, Sami Jokela
Helsinki University of Technology
TAI Research Centre
P.O. Box 9555, 02015 TKK, Finland
{eerika.savia, teppo.kurki, sami.jokela}@hut.fi

Abstract

The growing amount of new information has created a need for information filtering. Filtering is usually done on the textual content of the documents. Recent developments in the field of metadata suggest that filtering could be done according to only metadata (description of the actual document content).

In information filtering documents are matched against user interest profiles. This is based on some measure for similarity or distance between a representation of documents and user profiles. Both the representation and the distance measure should make comparisons meaningful. A common problem is caused by closely related concepts that are considered independent in the representation model. Furthermore, the matching should not be considered symmetric, since documents may cover some area of interest very well and should be matched against parts of the user profile instead of the whole profile.

In this paper we suggest a hierachical representation for describing documents and user profiles that attempts to model the related concepts. Our model includes an asymmetric distance measure that can also detect documents that cover some subtopic of an interest profile.

Keywords: Information filtering, Profiles, Personalisation, Similarity Measures, Metadata, Knowledge Discovery and Data Mining, World Wide Web

1. Introduction

Each day a rapidly increasing amount of new information is published on the Internet in a variety of different formats. In order to manage such an overflow, information filtering is required. However, if filtering is used to process the actual document contents as they are, bandwidth and performance difficulties are likely to appear. Therefore, we need a more compact representation of information content. This can be achieved by attaching metadata, i.e. information about information, to each document. Metadata can be used to store a description of the content of a multimedia document as well as the associated bibliographical information.

Automated information filtering requires also knowledge about the users. This knowledge can be represented in a user interest profile. Document metadata and user interest profiles should have compatible representations to make meaningful comparisons possible (Figure 1).

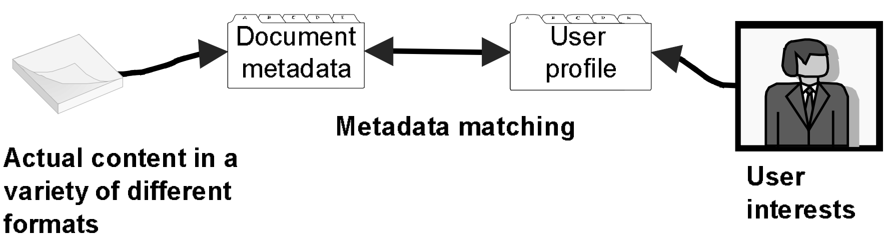

Figure 1 Document metadata and its relation to content and user interests

To manage the challenges described above, we have developed a hierarchical representation of the document content and user profile information. Utilising this representation document metadata and user profiles are matched to provide each user with a personalised information feed. Matching process is based on a distance measure that ranks a set of documents against a given profile and results in an approximation of the user's interest for each document. Our paper discusses different aspects of matching in this environment and the essential characteristics of this type of distance measures.

2. Metadata

Metadata is a critical component of an effective information management. Metadata has traditionally meant "data about data" or "information about information", but recently it has been suggested that it should mean "machine understandable information"[1]. Although a number of different proposals for metadata standardisation exist and more of them are underway (Dublin Core [2], RDF [3], XML [4]), this area still requires a substantial amount of work. An introduction to metadata standards can be found e.g. in [5], [6].

The challenges of creating descriptive metadata structures as well as the creation of the metadata itself are beyond the scope of this paper. We do not rely on any particular format but instead assume that suitable standardised structures for metadata exist.

In this paper we use a more restricted definition for metadata than some other sources do [7]. We use metadata to describe various aspects of the document content, such as subject matter, keywords, categorisation, location, story type and bibliographical information. Thus metadata contains a compact representation of the document and can be used to make filtering decisions without access to the original document. Once metadata is extracted from the actual content, it should be possible to transfer and process it independently and separately from the original content. This allows us to operate only on the metadata instead of the whole content. Metadata must be machine-readable, which means it must be standardised and described so that it can be processed without human intervention. This does not require the creation process to be fully automated, but once metadata is created, it can be interpreted and processed without human assistance.

We need uniform or compatible metadata structures to utilise content from different sources. In order to describe content effectively, metadata must be capable of supporting structured data with different types of values. Its format must also be flexible and expressive to accommodate different needs, for example to describe different media formats like video and audio as well as material in several languages (Figure 2).

In addition to a common format we need shared semantics for the metadata. This calls for a common vocabulary or ontology with which to describe different types of content. If content providers do not produce metadata in compatible metadata formats using shared or compatible ontologies the documents cannot be compared.

The information in metadata can be expressed using different data types. A document may be assigned to only one of several categories (a value from a discrete set of possibilities, e.g. media type in Figure 2) or the document may be assigned to several categories at the same time (a binary valued vector). A more expressive way would be a discrete distribution (a vector) with values from a continuous range (e.g. assigning weights for different keywords). Some properties can be represented by a value from a continuous range (e.g. time) or in principal even with a continuous distribution[1]. Also more complicated data structures, such as hierarchies or networks, can be represented in digital format by using a suitable, shared encoding scheme in interchanging metadata.

2.1 Dimensions of Metadata

In this paper the term *dimension* is used to describe different aspects of the content. We treat each dimension so that its concepts are independent of the concepts in the other dimensions. Independent dimensions could be for example concepts of generic news (subject matter -dimension in Figure 2), the media type of the document (Figure 2) and the time period relevant to the news article. The presence of several dimensions allows filtering according to different aspects.

---

[1] In digital format these are all either single discrete values or vectors with discrete values.

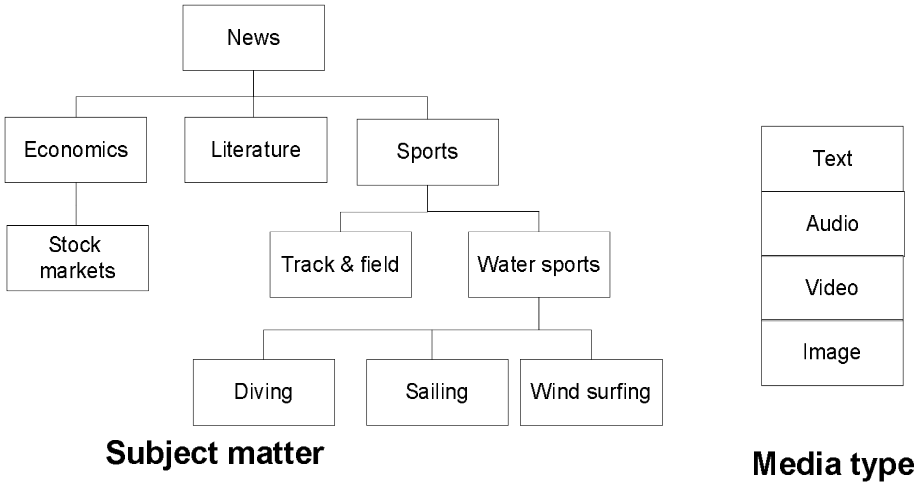

Figure 2 Examples of different dimensions of metadata and their structures in a news document. The subject matter of news content is described with a hierarchical structure. Media type is presented with one discrete value from a set of 4 possibilities.

3. Representing User Interests as Profiles

If we want to filter information to give a user a personalised information feed we need to represent his or her interests in digital format. We call this representation a *user profile*. When used in this context, user profile does not contain detailed information about the user's preferences for a certain software but general information about the user's interests in the filtering domain. In our application the main purpose of the user profile is to filter and rank documents so that the results reflect the actual interests of the user.

A user profile cannot represent the user's true interests in minute detail as the inner workings of the human mind still escape researchers. The representation has to be reasonably compact in terms of memory and computing power to allow processing on current computers. The processing has to take place in seconds or minutes depending on the type of the information feed that is filtered.

The user profile represents a mapping of the true interest profile to a more compact model space (Figure 3). The space is necessarily abstract, which means that the user profile is an approximation of the user's true real world interests in the representation space of the model.

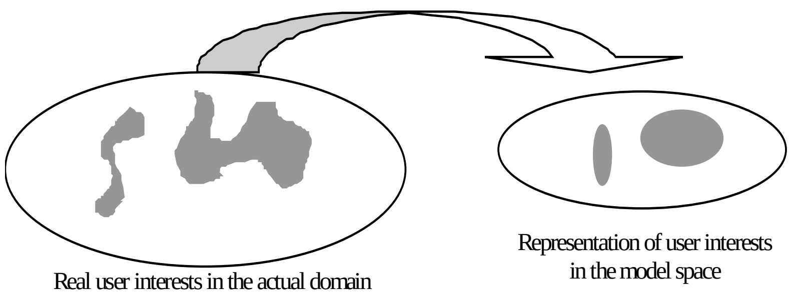

Figure 3 Profile as a representation of user interests in the model space

Most people's interests vary as time goes by and need for information on different topics fluctuates. Also the information feed to be filtered varies. Therefore, the user profile representation must be able to adapt gradually to the changes in the user's actual interests. Adaptation can be based on either implicit feedback (e.g. the user followed a suggested link) or explicit feedback (the user rates incoming documents). Adaptation to feedback in our system is discussed in [8].

Normally, a person's interests fall into several distinct groups. A user profile should be capable of representing different areas of interest and accurately match and rank documents accordingly. One way to represent distinct interests is to assign the user several independent user profiles, but this may be confusing or troublesome for the user.

A user's interests can be either very definite (for example related to a certain brand name) or very vague (new developments in artificial intelligence). Well-defined interests are easy to represent as keyword rules, but vague interest areas are better handled by generalising and grouping related concepts together. This gives rise to a generalisation hierarchy in the profile representation [9].

A user profile may also represent dislike for or avoidance of certain subject areas. This type of negative profile is discussed in [8] and is beyond the scope of this paper.

4. Matching

A crucial part of a filtering system is the matching of user profiles and documents. It has to be efficient and provide the users with a personally filtered information feed.

To be able to compare documents and profiles, we must have a way to measure how similar or different they are. What we mean by distance measure is a measure between a profile and a document that produces a non-negative real number reflecting the amount by which they differ from each other. It does not generally satisfy the requirements of distance measure as defined in mathematics (Appendix).

Although there have been many proposals for profile matching systems in literature, matching is usually done essentially in the same way. The measures in common use seem to be symmetric, which is not an eligible feature. The most popular information filtering method is the so-called vector space method [10], [11], [12].

4.1 Vector Space Model

In the vector space model, both documents and profiles are represented as vectors with components for different terms (*term vectors*). These components are weights that reflect the frequency of each term in the document and interest in a given term in the profile, respectively. The length of the resulting vectors can rise to tens or even hundreds of thousands.

The weight for a given term depends on the frequency of that term in the specific document compared to the frequency of that term in the whole document collection. Term frequency tf_{ik} expresses the strength of a given term in a given document. Inverse document frequency idf_k expresses the unusualness of a given term, thus emphasising terms with discriminating power. The weight of term k in document i is calculated

$w_{ik} = tf_{ik} \cdot idf_k$, and the resulting vector is finally normalised so that |d|=1.[2]

A commonly used similarity measure for these vectors is the *cosine measure* [11], [12]

$$\cos(u,d) = \frac{u \cdot d}{|u| \cdot |d|} = \frac{\sum_k u_k \cdot d_k}{\sqrt{\sum_k u_k^2}\sqrt{\sum_k d_k^2}} \in [-1,1],$$

which, of course, is symmetric. It measures the similarity of two vectors and can be used to measure their distance by using $dist(u,d) = 1 - \cos(u,d)$ instead.

For convenience, the profile vectors are also normalised so that |u|=|d|=1, which means that both document and profile vectors reside on the unit sphere [3]. The cosine measure treats both the document and the profile as vectors of $\mathbf{R}^n$ and examines the angle between those vectors. It thus assumes the properties of vector space $\mathbf{R}^n$, including orthogonality of the vector components.

Adaptation of profile in the vector space model can be implemented by adjusting the profile vector towards or away from feedback document's vector [11].

---

[2] Euclidean norm $|x| = \sqrt{\sum_k x_k^2}$

[3] Actually, the document vectors are restricted to that segment of the unit sphere where all the components are non-negative. Profiles need not be positive.

4.2 Orthogonality and Latent Semantic Indexing

The adopted methods for calculating similarity or distance between user profiles and documents have typically certain restrictions which are not considered in most cases. For cosine measure to be meaningful the components of the vectors should be orthogonal. That is usually not the case, because different terms may depend on each other.

What happens if the components are not orthogonal? Assume we have only seven terms in our vector space: *Literature*, *Economics*, *Artificial intelligence*, *Diving*, *Swimming*, *Windsurfing*, and *Water Skiing*.

If we take a closer look at these concepts we can see strong dependency between the four concepts related to water sports. We could categorise them into one concept, resulting in four approximately orthogonal terms, namely *Literature*, *Economics*, *Artificial Intelligence* and *Water Sports* (containing *Diving*, *Swimming*, *Windsurfing*, and *Water Skiing*)

7 concepts	u	doc1	doc2
AI	0.6		
Economics			$\sqrt{3} \cdot 0.4$
Literature			$\sqrt{3} \cdot 0.4$
Diving	0.8	0.2	0.2
Swimming		$\sqrt{2} \cdot 0.4$	
Surfing		$\sqrt{2} \cdot 0.4$	
Water skiing		$\sqrt{2} \cdot 0.4$	

Table 1

4 concepts	u	doc1	doc2
AI	0.6		
Economics			$\sqrt{3} \cdot 0.4$
Literature			$\sqrt{3} \cdot 0.4$
Water Sports	0.8	1.0	0.2

Table 2

The example in Table 1 shows one user profile *u* and documents *doc1* and *doc2* [4]. If we compare the documents to the user profile using the cosine measure and all 7 terms we get the following results:

cos(u,doc1) = 0.16

cos(u,doc2) = 0.16 .

When applied to the uncategorised terms, the cosine measure implies no difference between the results, even though the profile of *doc1* contains concepts closer to person's interests than *doc2* and thus probably reflects the person's interests better. If we consolidate the water sports into one term (Table 2), the calculation yields

cos(u,doc1) = 0.8

cos(u,doc2) = 0.16 ,

which makes the documents very different in regard to the profile. Thus relations between concepts contradict the orthogonality assumption of the cosine measure.

Remarkable work in solving the problem of non-orthogonality of terms has been done by Foltz and Dumais [13]. They have developed a method called Latent Semantic Indexing (LSI), which is based on methods of linear algebra, singular value decomposition (SVD) to be exact. LSI can be applied to any fixed document collection that is represented in term vectors. By decomposing the matrix that consists of the document vectors, one gets an orthonormal basis for that document space, and if the calculations are done according to that basis, the components are indeed orthogonal. In addition, the least significant basis vectors are removed in order to reduce the high dimension of the space.

One problem of this method is the intensive calculation needed to compute the SVD and, of course, finding a good enough collection of documents to represent all the future documents as well. The same problem of finding the collection arises also in vector space method, if one would like to have the inverse document frequencies computed (Appendix). The resulting basis vectors are of length the number of terms (does not differ from vector space method here) and have a certain weight for every term. Thus they cannot be feasibly handled or comprehended by humans and the calculations have to be performed automatically with no

[4] normalised in Euclidean norm

intuitive presentation to the user. Our aim is to develop a document classification system that is understandable to humans and is not restricted to text documents, and thus we prefer a hierarchical representation to LSI [9].

4.3 Hierarchical Metadata and Asymmetric Distance Measure

It is obvious that the suggested information retrieval methods [10] only apply to text documents, but one seldom mentioned drawback is their language dependency. For every language one needs a separate term list, document collection and handling. Another problem is word stemming, which requires the use of language analysis tools in many languages, including Finnish.

Both problems can be solved by generating metadata in a suitable and unified format for all different media types. If we can spend extra effort to create metadata that accurately and expressively reflects the content of a multimedia document, we can use the created metadata for matching. Furthermore, using structured metadata and a shared ontology to describe the multimedia content we can create new, more meaningful ways for matching user profiles and multimedia documents.

Once the ontology is created we can use it to model both document content and user profiles. We claim that symmetric distance measures cannot catch the interest distribution. Although the syntactic form of the profile and document metadata are the same their semantics are not exactly the same. A person's interest distribution usually consists of many separate interests and the best matching documents may not be the ones that try to cover all the interests at the same time. For example, if a person's interest profile consists of certain train timetables, AI research and apartment sales advertisements, it would be peculiar to assume the best matching documents to be the ones with a touch of each subject. Therefore, the distance measure should not be symmetric. What we try to do is match the documents onto parts of the user profile. The role of the profile then becomes distinct from the role of the document and the measure is no longer symmetric.

Let us examine documents *doc1* and *doc2* in Table 3 shown also in Figure 4[5]. Both documents intersect with the profile only in the topic of Artificial Intelligence and neither one of them matches the other interests the user has. But there is a difference between these two cases: *doc2* is related only to the topics of the profile while *doc1* involves Economics, which is not present in the profile at all. *Doc2* does not involve topics outside the profile, so it can be seen as a perfect match to a part of the user profile. We can conclude that the white area in Figure 4 is not as significant as the striped area when evaluating the distance.

Concept	User profile	doc1	doc2	doc3
AI	**0.2**	**0.4**	**1.0**	
Train time tables	**0.3**			
Apartment sales advertisements	**0.5**			**0.4**
Economics		**0.6**		**0.6**

Table 3

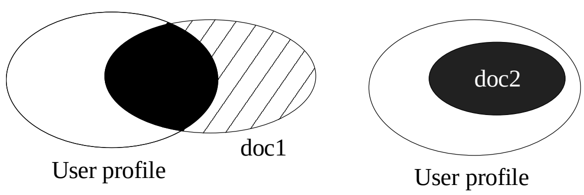

Figure 4

We suggest that a hierarchical structure could be used in producing and storing content description metadata. For a human hierarchy is an intuitive way to navigate a multitude of terms. Hierarchical structure does not

[5] The sum of the weights in each document is normalised to 1.

satisfy the requirement of orthogonality, but one can naturally choose the distance measure in a more suitable way. It has also the advantage of not requiring heavy calculations. The hierarchy itself codifies information about the dependencies of different terms. Distance calculations can be done on different levels of the hierarchy and the topmost level can be seen as being approximately orthogonal. The topmost level can efficiently reduce the amount of required calculations, since matching is needed only for those profile-document pairs that have a lot in common. We need to consider only the relevant parts of the hierarchy instead of the whole term vector with lots of zero entries. This approach scales well as the number of documents and profiles grow.

Figure 5

Furthermore, we suggest a fuzzy representation in the metadata hierarchy that describes document content [8], [14]. It can be seen as a discrete distribution on the leaf nodes of the concept tree. The distribution is normalised so that the sum of weights equals 1. The advantage of this is that it has more expressive power than just crisp classification of documents into separate classes. The upper levels of hierarchy simply gather related subjects together and the weight of upper nodes is always the sum of the weight of its child nodes (Figure 5). Since the normalisation is done in the 1-norm (Appendix), it makes sense to compare the distributions (document and profile) also according to 1-norm. The distance calculations can be performed on any level of the hierarchy and the results of different levels can be combined, e.g., by taking a weighted sum over them.

Our asymmetric measure *e(u,d)* for each level of hierarchy is

$$e(u,d) = \sum_h d_h \qquad \in [0,1] \,,$$

where d_h are those document weights where the corresponding weight in user profile *u* is zero ($u_h = 0$). This means that we only sum over the document weights where the profile weight is zero. We call this measure the *coarse measure*, since it gives a rough estimate for the distance between the document and the profile. This measure only takes into account the amount by which the document misses the profile (the striped area in Figure 4). It would treat documents *doc1* and *doc3* in Table 3 as equally well matching documents.

Another measure that covers the area where both the profile and the document have weight is needed (the black intersection area in Figure 4). Our proposal for each level of hierarchy is

$$f(u,d) = 1 - \sum_k u_k \cdot d_k \qquad \in [0,1] \,,$$

which is symmetric. We call it the *fine measure*[6].

---

[6] It is not the cosine measure though it looks like it. *u* and *d* are normalised according to 1-norm.

The question that arises is how to combine these two measures to get one unified measure *D(u,d)* that ranks the documents for the profile. For example, the interval [0,1] can be divided into *n* subintervals[7] of length *h* and the coarse measure *e(u,d)* can be truncated to the lower bound of the subinterval. Those documents that are ranked equally in this phase are then further ordered by the fine measure *f(u,d)*.

$$D(u,d) = h \cdot \left\lfloor \frac{1}{h} \cdot e(u,d) \right\rfloor + h \cdot f(u,d) \ ,$$

where $\lfloor \cdot \rfloor$ denotes the floor function. This measure takes into account both the intersecting region in black and the striped area in Figure 4 but ignores the white region, as was desirable.

A more detailed representation of this model and comparison to other distance measures can be found in [14].

5. Conclusions and Future Work

We conclude that the most frequently used distance measures are very demanding because they rely on the assumption that the concepts are orthogonal, that is, fully independent of each other. Either concepts should be made orthogonal (LSI) [13] or distance measure should take the dependency into account, or even both. Furthermore, the meaning of a user profile is different from the meaning of a document representation (metadata). Thus matching should be done asymmetrically to catch parts of the overall user interest profile.

To meet these requirements we have developed a hierarchical representation for document subject matter and an asymmetric distance measure. Our approach rests on the assumptions that there is content description metadata available for the filtering process and that the metadata is based on a sufficiently expressive ontology.

So far we have been working on the proof of concept implementation and testing on artificial material. We are currently setting up real world cases to test the new approach with actual material. Our methods also need further consideration. Adaptation of profiles based on feedback is covered in [8] but needs further development and analysis. Our future work also includes taking negative user profiles into account in the matching process. Different ways to combine the coarse and the fine measure have to be still examined. Exploring new candidates for the fine measure *f(u,d)* seems necessary, since the measure presented here has not been under thorough examination and it is presumably very sensitive to the number of intersecting nodes.

[7] The appropriate value for parameter n must be found experimentally.

6. Acknowledgements

This research is part of the SmartPush project sponsored by the Finnish Technology Development Centre TEKES and Alma Media, WSOY, Sonera, ICL, Nokia Research Centre and TeamWARE Group.

7. References

[1] W3Ca (1998). Metadata Activity, W3C Technology and Society domain, http://www.w3.org/Metadata/Activity.html

[2] Weibel, S. & Miller, E. (1997). Dublin Core Metadata Element Set WWW homepage, http://purl.org/metadata/dublin_core

[3] Lassila, O. (1997). Introduction to RDF Metadata, W3C NOTE 1997-11-13, http://www.w3.org/TR/NOTE-rdf-simple-intro

[4] W3Cc (1998). Extensible Markup Language (XML™), W3C Architecture Domain, http://www.w3.org/XML/

[5] W3Cb (1998). Metadata and Resource Description, W3C Technology and Society domain, http://www.w3.org/MetadataW3Cb, URL: http://www.w3.org/Metadata

[6] IFLA (1998) Metadata Resources, http://www.nlc-bnc.ca/ifla/II/metadata.htm

[7] Berners-Lee, T. (1997). Metadata Architecture, http://www.w3.org/designIssues/Metadata

[8] Savia, Eerika (1997). Adaptiivinen oppimismenetelmä käyttäjäprofiilien päivittämiseen. Matematiikan erikoistyö, TKK. http://smartpush.cs.hut.fi/pubdocs/

[9] Bloedorn, E., Mani, I., MacMillan R. (1996). Representational Issues in Machine Learning of User Profiles. AAAI Spring Symposium on Machine Learning in Information Access.

[10] Salton, G., McGill, M.J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill

[11] Sheth, Beerud (1994). A Learning Approach to Personalised Information Filtering, M.I.T. http://agents.www.media.mit.edu/groups/agents/papers/

[12] Yan, Tak W.,Garcia-Molina, Hector (1994). Index Structures for Information Filtering under the Vector Space Model. IEEE Conference on Data Engineering.

[13] Foltz, P.W., Dumais, S. (1992). Personalised Information Delivery : An Analysis of Information Filtering Methods. Communications of the ACM 35(12)) http://www--psych.nmsu.edu/~pfoltz/cacm/cacm.html

[14] Savia, Eerika (1998). Etäisyysmittojen vertailu elektronisessa täsmäjakeluympäristössä. Matematiikan erikoistyö, TKK. http://smartpush.cs.hut.fi/pubdocs/