AN ONTOLOGY- CONTENT-BASED FILTERING METHOD

Peretz Shoval, Veronica Maidel, Bracha Shapira

Abstract: Traditional content-based filtering methods usually utilize text extraction and classification techniques for building user profiles as well as for representations of contents, i.e. item profiles. These methods have some disadvantages e.g. mismatch between user profile terms and item profile terms, leading to low performance. Some of the disadvantages can be overcome by incorporating a common ontology which enables representing both the users' and the items' profiles with concepts taken from the same vocabulary.

We propose a new content-based method for filtering and ranking the relevancy of items for users, which utilizes a hierarchical ontology. The method measures the similarity of the user's profile to the items' profiles, considering the existing of mutual concepts in the two profiles, as well as the existence of "related" concepts, according to their position in the ontology. The proposed filtering algorithm computes the similarity between the users' profiles and the items' profiles, and rank-orders the relevant items according to their relevancy to each user. The method is being implemented in ePaper, a personalized electronic newspaper project, utilizing a hierarchical ontology designed specifically for classification of News items. It can, however, be utilized in other domains and extended to other ontologies.

Keywords: Ontology, Retrieval models, Information filtering, Content-based filtering, User profiles.

ACM Classification Keywords: H.3 Information Storage and Retrieval, H.3.1 Content Analysis and Indexing, H.3.3 Information Search and Retrieval, I.7 Document and Text Processing.

1. Introduction

In content-based filtering, the representations of the content of items (e.g. documents, News), i.e. the items' profiles, are compared with the representation of the users, i.e. the users' profiles. In order to enable matching and measuring the similarity between the profiles, it is assumed that a user's profile and an item's profile share a common method of representation (e.g., by keywords). The output of the matching process can be expressed as a ranking score, indicating the similarity between the user's profile and each item.

A user profile can be generated in various ways, including explicit definition by the user, or implicit analysis of the user's behavior (e.g. by logging and analyzing what the user read). An item's profile too can be generated in various ways, e.g. explicitly, by asking the originator (author) to specify proper index terms, or automatically, using a text classification algorithm which extracts terms representing the item's content in the best way. At any rate, no matter which method is used for creating either type of profile, content-based filtering has drawbacks due to well known problems of term ambiguity. For example, different terms may be used to represent the same content or the same user (synonymy); or, the same term may be used to represent different contents or different users (homonymy).

A possible way to overcome such problems of ambiguity might be through the use of ontology, i.e., a controlled vocabulary of terms or concepts, and semantic relationships among them. Ontology can bridge the gap between the terms in the users' profile and the terms representing the items. Ontology can be organized in various ways. For example, a taxonomy is a hierarchical structure with is-a relationships; in a thesaurus there are a few more types of relationships, e.g. BT/NT (broader-term; narrower terms) and general relatedness. Note that a thesaurus is a graph, not a hierarchy, because a term may have many NTs and more than one BT. In the newspapers domain, which is exemplified in this study, there is an ontology specifically generated for classification of News named NewsCodes, created by IPTC [Le Meur and Steidl, 2004].

Assuming that there exists ontology of a specific domain, which is used for representing users (user profiles), and contents of items (item profiles), the research question we deal with is how exactly to match and measure the similarity between a user's profile and an items' profile. Obviously, if a user's profile includes exactly the same concept (terms) as an item's profile, there is some similarity between them; but the two profiles may include

different concepts and still be similar to a certain degree – depending on if and how "close" the concepts are in the two profiles with respect to the common ontology.

This research is conducted within the framework of *ePaper*, an electronic personalized newspaper research project, which is aimed to provide a personalized electronic newspaper to each reader. In the News domain, instant filtering of News items is important. Since a new item has no reading history, the filtering and personalization cannot rely on collaborative filtering (as opposed to other domains such as recommendation of books, movies, etc.), but rather need to rely on content-based filtering, so that once a new item arrives to the News repository, the content-based filtering algorithm can perform the necessary matching with the users' profiles and determine the degree of relevancy of each item to the potential users. If many News items accumulate in a certain period of time, the content-based filtering algorithm can rank-order the items according to their relevancy to each of the potential users.

The remaining of this paper is structured as follows: The next section provides a background on content-based filtering and on ontological modeling, and reviews related research on conceptual and ontological modeling employed in content-based filtering. Section 3, the main section of the paper, presents the proposed method for the ontology- content-based filtering, along with an example. Section 4 describes the evaluations that we plan to conduct with the proposed method, and the last section summarizes and proposes further research and extensions to the proposed method.

2. Background on Content-Based Filtering and Ontological Modeling

2.1. Content-Based Filtering

The information filtering approach is based on the information retrieval (IR) domain and employs many of the same techniques [Hanani et al., 2001]. One aspect by which information filtering differs from IR is with respect to the uses' interests. While in IR the users poses ad-hoc queries, in information filtering the users have profiles which represent their long-term interests, and the filtering system tries to provide to each user relevant items on a long-term basis. As said, the user profiles, as well as the item profiles, may consist of sets of terms. Based on some measure of similarity between the respective profiles, the filtering system selects and rank-orders the relevant items and provides them to the user.

The actual relevancy of an item provided by the system to a user can be determined by explicit or implicit user feedback. Explicit feedback requires the user to express the degree of relevancy of the provided item, while in implicit feedback the relevancy of the item is inferred by observing the user's behavior, e.g. the reading time. Implicit feedback may be more convenient for the user but more difficult to implement and less accurate. User feedback enables to update the user's profile according to what he/she actually read, liked or disliked.

There exist two main approaches in information filtering: collaborative and content-based. In collaborative filtering, the system selects and rank-orders items for a user based on the similarity of the user to other users who read/liked similar items in the past. In content-based filtering, the system selects and rank-orders items based on the similarity of the user's profile and the items' profiles.

A major advantage of content-based filtering is that users can get insight into the motivation why items are considered relevant to them, because the content of each item is known from its representation. Content-based filters are less affected by problems of collaborative filtering systems [Claypool et al., 1999] such as "cold start" and sparsity: if a new item is added to the repository, it cannot be recommended to a user by a collaborative filter before enough users read/rated it. Moreover, if the number of users is small relatively to the volume of items in the repository, there is a risk of the ratings coverage becoming very sparse, thinning the collection of recommendable items [Balabanovic and Shoham, 1997]. For a user whose tastes are unusual compared to the rest of the population, the system will not be able to locate users who are particularly similar, leading to poor recommendations.

However, content-based filtering has disadvantages too, for example the fact that it focuses on keyword similarity. This approach is incapable of capturing more complex relationships at a deeper semantic level, based on different types of attributes associated with structured objects of the text [Dai and Mobasher, 2001]. Consequently, many items are missed and many irrelevant items are retrieved [Blair and Maron, 1985].

Unlike humans, content-based techniques have difficulty in distinguishing between high quality and low quality information, since both good and bad information might be represented by the same terms. As the number of items increases, the number of items in the same content-based category increases too, further decreasing the effectiveness of content-based approaches [Claypool et al., 1999]. Another disadvantage of content-based methods is that they require analyzing the content of the document, which is computationally expensive and even impossible to perform on multimedia items which do not contain text.

To expand the first point of the disadvantages, it can be added that there is a tremendous diversity in the words that people use to describe the same concept (synonymy) and this places strict and low limits on the expected performance of keyword-based systems. If the user uses different words from the organizer (indexer) of the information, relevant materials might be missed. On the other hand, the same word can have more than one meaning (homonyms), leading to irrelevant materials being retrieved [Dumais et al., 1988]. This disadvantage is added to the fact that the basic models of content-based filtering assume a representation of documents as sets or vectors of index-terms and typically employ only primitive search strategies based solely on the occurrence of term or combinations of terms [Knappe, 2005].

Thus, extensions to the traditional content-based filtering methods should be considered. Extensions may include additional knowledge in the form of a coherent taxonomy of concepts in the domain spanned by the items. This type of conceptual knowledge would provide means for item and user profile representation.

There is a need for devising a content-based approach which extends the classical models, where the use of simple natural language analysis in combination with the knowledge contained in an ontology forms the basis for representations of both user profiles and items. Consequently, items can be described using a concept language and be directly mapped into the ontology. The similarity between such representation of the user and the representations of the items will be based on the proximity principle stating that the distance of two descriptive items in the ontology is directly related to their similarity [Knappe, 2005].

2.2. Ontological Modeling

Ontology is a specification of a conceptualization. It can be described by defining a set of representational concepts. These definitions are used to associate the names of entities in the universe (e.g., classes, relations, functions or other objects) with human-readable text, describing what the names mean, and formal axioms that constrain the interpretation and focus the well-formed use of these concepts [Khan, 2000]. When constructing ontology, not only concepts and relationships are defined, but also the context in which the concept (relationship) applies. Therefore, ontology defines a set of representational terms which are called concepts, and the interrelationships among the concepts.

Linguistic ontologies (e.g., WordNet) and thesauri express various relationships between concepts (e.g. synonyms, antonyms, is-a, contains-a), and have a hierarchical structure based on the relations between concepts. But they do not explicitly and formally describe what a concept means [Khan, 2000]. WordNet, for example, is an electronic lexical database that contains nouns, verbs, adjectives and adverbs which are organized into synonym sets (*synsets*), each representing one underlying lexical concept [Magnini and Strapparava, 2001]. It is offering two distinct services: a vocabulary which describes the various word senses, and an ontology which describes the semantic relationships among senses [Guarino et al., 1999].

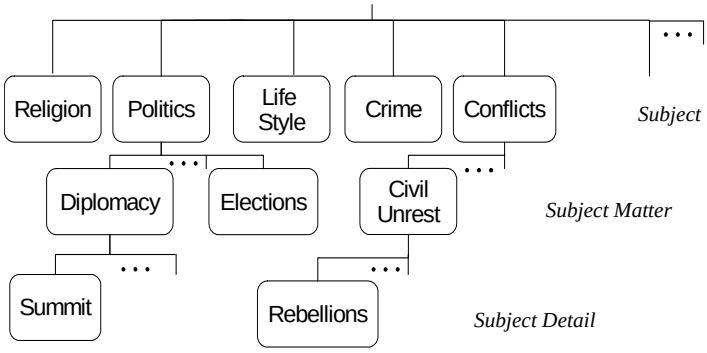

Figure 1: Example of IPTC NewsCodes ontology

An example of domain ontology is the IPTC NewsCodes [Le Meur and Steidl, 2004]. This is a 3-level hierarchical ontology of concepts targeted to News description; it currently contains approximately 1,400 concepts. A first level concept of NewsCodes is called Subject; a second level – Subject Matter, and a third – Subject Detail. Figure 1 demonstrates an example.

2.3. Related Work

Savia et al. [1998] was one of the first to present a hierarchical representation for describing documents and user profiles by attaching metadata to each document and using the same method to generate a compatible representation of users' interests. A hierarchical representation was chosen in order to develop a document classification system understandable to humans and yet not restricted to text documents. Savia et al. chose to take advantage of the hierarchical metadata concepts along with an asymmetric distance measure, which considers not only the concepts appearing both in the user's profile and in the document's profile, but also concepts appearing in the document's profile which do not appear in the user's profile. The underlying assumption for the asymmetric measure was that the best matching documents are not necessarily those that cover all the interests at the same time. Distance computations were performed on different levels of the hierarchy and the metadata was represented in a fuzzy distribution among the leaf nodes of the concept tree.

Ontological and conceptual modeling was used in order to extract user profiles, such as the four-level ontology used in the Quickstep system [Middleton et al., 2001] which recommends papers to researchers by combining both content-based and collaborative filtering techniques. Papers were represented as term vectors with term frequency normalized by the total number of terms used for a term's weight. Whenever a research paper was browsed and had a classified topic, it accumulated an interest score of that topic for the particular user. In the ontology-based user profile, whenever a topic received some interest all its super classes gained a share: the immediate super-class gained 50% of the main topics value; the next super-class gained 25%, and so on. This way, general topics rather than just the most specific ones were also included in the profile and thus produced a broader profile. Recommendations were computed based on the correlation between the user's topics of interest and papers classified to those topics.

Another work which used ontology for content-based retrieval was the electronic publishing system CoMet [Puustjärvi and Yli-Koivisto, 2001]. CoMet extracted metadata information both about users and about contents of documents (document profiles) and stored the metadata in hierarchical ontology structures. Comparison between a user's profile and the documents' profiles was performed by finding the largest combined hierarchy (LCH), which is the largest hierarchy that the user profile and the document profile share in the ontology. By using the weights on the nodes in each level, a similarity measure was calculated between the documents that had an LCH with a user profile. In a weighted LCH-matching, the deepness of the LCH was emphasized in the matching calculations, since the depth of the hierarchy has a significant effect on the expression power of the incorporated ontology. The depth of the profile was also suggested to be used as a generalization tool. For example, if a user is interested in news items on F1 (a sport car), one can assume that she would like to view other motor sport related items when F1 news items are not available. The result of the matching generated a set of news items most suitable for the user according to the calculation result of LCH-matching.

Pereira and Tettamanzi [2006] illustrated a novel approach to learning users' interests on the basis of a fuzzy conceptual representation of documents, by using information contained in ontology. Instead of a keyword representation, documents were represented as a vector of components expressing the “importance” of the concepts. In order to choose the concepts that would represent a document, they considered both the leaf concepts and the internal nodes of the ontology. The internal nodes were implicitly represented in the importance vector by “distributing” their importance to all their descendants down to the leaf concepts. All documents with a certain level of similarity were grouped together into fuzzy clusters, in order to express user interests with respect to clusters instead of individual documents. Since the clusters were fuzzy, each document received its membership degree for that cluster, meaning it could belong to more than one cluster. A user model was represented as a vector of membership degrees which described the model’s guess of the extent to which the user was interested in each document cluster. A user profile was set up by adding to the list of its interest groups the instances of clusters with features similar to those requested by the user.

In the above survey we have emphasized methods involving the incorporation of ontologies both for user profile generation and for representation of items. Some of the methods employed ontology in order to acquire user profiles more accurately, while others used ontology in order to perform disambiguation of a user profile. In most cases, the ontology was used in all of the steps taken towards the retrieval of items according to the user profile. All studies which incorporated ontology in their content-based filtering method provided better and more accurate results compared to traditional content-based methods. This encouraged us to adopt the ontology approach and inspired us to introduce a novel filtering method which incorporates ontology.

3. The New Method for Ontological-Content-based Filtering

3.1. Research Goal

The aim of this research is to develop, implement and evaluate a new ontology-based filtering method, which filters and ranks relevant items by measuring the similarity of user profiles and item profiles, both consisting of ontology concepts, by considering the "closeness" (or distance) of concepts in the profiles, based on their location in the ontology. We utilize the method in the News domain, as part of *ePaper*, a research project which includes the development of a personalized electronic newspaper system. In this research we incorporate ontology for the News domain and exploit its three-level hierarchy in the representation of user profiles and News items profiles, and in the process of matching between them.

3.2. The Ontological-Content-based Filtering Method

The filtering method, initially proposed by Shoval [2006], is based on the assumption that each item (e.g. a News item) and each user profile (e.g. a reader of the *ePaper*) is represented with a set of concepts taken from the ontology. In the *ePaper* system we use the IPTC NewsCodes ontology, which is exemplified in Figure 1. It may be assumed that the generation of an item's representation (profile) is done automatically, utilizing some classification technique which analyses both the metadata describing the item and the actual text of the item. (We do not elaborate here on how this is done because as it is not an essential part of the proposed method.) Similarly, it may be assumed that an initial user profile is generated explicitly by the user who selects concepts from the ontology and assigns them weights of importance. Subsequently, the concepts in the initial user profile and their weights are updated implicitly, based on monitoring the items actually read by the user and considering the ontology concepts by which those items are represented. (This part is not elaborated here as well; suffice is to know that at any point in time a user's profile contains an up-to-date weighted set of ontology concepts.)

Following are the details of the filtering method.

Representation of contents – an item's profile:

An item's profile consists of a set of ontology concepts which represent its content. The concepts representing an item are the most specific ones in a certain branch of the hierarchy. For example, if an item deals with 'sport' and specifically with 'football', it is represented with 'football' concept only; the ontology can tell that the latter is a child (subtype) of the former.

Obviously, an item may be represented with many ontology concepts; each concept may appear in any branch of the ontology hierarchy and at any level – all depending on the actual content of that item. For example, an item's profile may include the concepts 'politics' (a top-level concept), 'football' (child of 'sport') and 'rebellions' (grandchild of 'conflicts'). Note that the profile may include sibling concepts, i.e. children of the same super concept. For example, an item's profile may include both 'football' and 'basketball' (children of 'sport').

Note that we do not assume that the concepts representing an item are weighted, although the proposed filtering algorithm can be adjusted for such possibility.

Representation of users – a user's profile:

A user's content-based profile consists of a weighted list of ontology concepts representing his/her interests. Obviously, a user's profile may consist of many ontology concepts, each appearing in different branches and at different levels of the hierarchy. For example, a user's profile may include the concept 'sport' only, or 'sport' and 'football', or 'football' and 'basketball', or all the three – besides many other concepts. This means that a certain concept in an item's profile may be "matched" (i.e. compared) with more than one equivalent concept in the user's profile. For example, if an item's profile includes 'football' and a user's profile includes both 'sport' and 'football', then there is a "perfect match" between the two profiles due to the common concept 'football', and also a "partial match" due to the parent concept 'sport'.

As stated before, the user's content-based profile may be generated initially by the user who selects concepts from the ontology and assign them weights of importance. (The total of the weights is normalized 100%). Then, the user's profile is constantly updated according to implicit feedback from the user: when a user reads an item and finds it interesting, the concepts in that item's profile which are not yet in the user's profile are added to it, and the weights of all concepts in this profile are recalculated as follows: a new concept is added with 1 'click' (a 'click' indicates how many times that concept was found relevant to the user) and the weight of an existing concepts is

increased by 1 'click'. The weight of each concept in the user's profile is the number of its 'clicks' divided by the total number of 'clicks' in the user's profile. (Hence, the weights sum up to 100 %.)

Measuring similarity between an item and a user:

An item and a user are similar to a certain degree if their profiles include common (the same) concepts or related concepts, i.e. concepts having some kind of parent-child relationship. An item's profile and a user's profile may have many common or related concepts; obviously, the more common or related concepts, the stronger is the similarity between them. For example, if a user's profile includes 'football' and 'sport', this profile is similar (to a certain degree) to an item including these two concepts, but it is less similar to an item including just 'sport', and is more similar to an item including 'sport', 'football' and 'basketball'.

In the *ePaper* project, we adopted the 3-level NewsCodes ontology, so related concepts may be only one or two levels apart (parent-child or grandparent-grandchild), but generally concepts may be more levels apart. It is obvious that the closer two concepts are in the ontology, the closer are the two objects which they represent (i.e. the user and the item).

When dealing with related concepts appearing a user's profile and in an item's profile, two different cases can be distinguished: in one case, the concept in the user's profile is more general than the related concept in the item's profile (one or two levels apart), meaning that the user has a more general interest in the topic which the item deals with. In the other case, the concept in the user's profile is more specific than the related concept in the item's profile (one or two levels apart), meaning that the user has more specific interests in the topic dealt in the item. In any of the above cases of "partial match" between the user and item concepts, we should also consider the distance between the concepts: two related concepts which are only one level apart are closer (i.e. more similar) than two concepts which are two levels apart.

Scores of similarity:

Based on the above, in a 3-level hierarchical ontology we can distinguish between 9 different possible cases of similarity between concepts in a user's profile and an item's profile, as portrayed in Figure 2.

- "Perfect match": the concept appears both in the user's profile and in the item's profile. I1, I2, I3 (see Figure 2) denote the level of a concept in an item's profile, and U1, U2, U3 - the level of a concept in the user's profile. A 'perfect match' can occur in 3 cases:
  - I1=U1 (e.g. both item and user profiles include 'sport')
  - I2=U2 (e.g. both item and user profiles include 'football')
  - I3=U3 (e.g. both item and user profiles include 'Mondeal games')

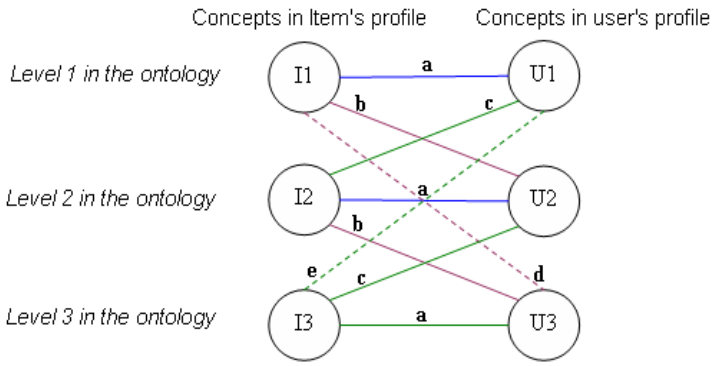

Figure 2: Hierarchical similarity measure

- "Close match": a concept appears only in one of profiles, while a parent or child of that concept appears in the other profile. A 'close match' can occur in 2 pairs of cases:
  - I1=U2 (e.g. item concept is 'sport', while user concept is 'football')
  - I2=U3 (e.g. item concept is 'football' while user concept is 'Mondeal games')

  In the above 2 cases, the item's concept is more general than the user's concept (1 level apart), i.e. the user interest is more precise/specific than the item.
  - I2=U1 (e.g. item concept is 'basketball' while user concept is 'sport')
  - I3=U2 (e.g. item concept is 'Euro league' while user concept is 'basketball')

  In the above 2 cases, the item concept is more specific than the user concept, i.e. the user's interest is more general.

Note that in all the above 4 cases there may be more than one occurrence of 'close match' between the concepts. For example, in the case I1=U2, assume the item's concept is 'sport' while the user's profile includes both 'football' and 'basketball' concepts. When measuring similarity, we have to consider all possible 'close matches' between parent and children concepts.

- "Weak match": a concept appears in one profile, while a grandparent concept or a grandchild concept appears in the other profile (concepts are 2 levels apart). A 'weak match' can occur in 2 cases:
  - I1=U3 (e.g. item concept is 'sport' while user concept is 'Mondeal games') – in this case the item is much more general than the user's interest.
  - I3=U1 (e.g. item concept is 'Euro league' while user concept is 'sport') – in this case the item is much more specific than the user's interest.

  Recall that there may be more than one occurrence of 'weak match' between the concepts. For example, in the case I3-U1 the user concept is 'sport' while the item concepts include 'Euro league' and 'Mondeal games'.

For each of the 9 possible cases we determine a score of similarity. In the 3 cases of "perfect" match' labeled 'a' (see Figure 2) the score is 1 (maximal); in all other cases the score should be less than 1, depending if it is a 'close' or a 'weak' match and on the "direction" of the relationship, i.e., whether the user's concept is more general or more specific than the item's concept. For example, the score for the case I1=U2 (the item's concept is more general than the user's concept) may be 2/5, while the case I2=U1 (the item's concept is more specific than the user's concept) may score 2/3 – higher. The rationale for this may be that in the first case the item deals with a more general concept than the user's interest, yielding lower Precision than in the other case, where the item deals with a more specific concept than the user's interest, thus yielding higher Precision. But this assumption, as well as the exact scores of similarity for all possible cases is subject to experimentation.

The following is a possible scoring scheme for the 9 possible cases:

- I1=U1 $\rightarrow$ 1; I2=U2 $\rightarrow$ 1; I3=U3 $\rightarrow$1 (3 cases of "perfect match"; marked a in Figure 2)
- I1=U2 $\rightarrow$ 2/5; I2=U3 $\rightarrow$ 2/5 (2 cases of "close match" - item concept is more general; marked b)
- I2=U1 $\rightarrow$ 2/3; I3=U2 $\rightarrow$ 2/3 (2 cases of "close match" - item concept is more specific; marked c)
- I1=U3 $\rightarrow$1/3 (case of "weak match" - item concept is much more general; marked d)
- I3=U1 $\rightarrow$ 1/2 (case of "weak match" - item concept is much more specific; marked e)

Measure of similarity between item and user:

The similarity of an item's profile to a user's profile is based on the number of "perfect match", "close match" and "weak match" of concepts between the two profiles, and on the weights of the concepts in the user's profile. The overall Item Similarity score (IS) is computed as follows:

$$IS = \frac{\sum_{i \in Z} N_i \cdot S_i}{\sum_{j \in U} N_j}$$

where:

Z - number of concepts in item's profile
U - number of concepts in user's profile
i - index of the concepts in item's profile
j - index of the concepts in user's profile
S_i - score of similarity, depending if it is a "perfect", "close" or a "weak" match of concept i to a respective concepts in the user's profile. (Note that in case of no match at all, $S_i = 0$.)
N_i - number of clicks on the concept (used to determine the concepts' weights)

The matching algorithm:

The algorithm can be applied for measuring the similarity of a single item to a single user, or for rank ordering by relevancy a batch of items for a single user, or for rank ordering by relevancy a batch of users for a single item, or for rank ordering by relevancy a batch of items for a batch of users – all depending on the specific need/application.

The algorithm described below is applied for measuring the similarity of a single item's profile to a single user's profile. The algorithm is expressed in pseudo-code; it does not refer to any specific programming language,

database system and other implementation aspects. However, it may be assumed that due to size on one hand and efficiency on the other hand, during execution the ontology resides in memory.

Since a user's profile may include many concepts (depending, among else, on how many items he already read), some with very low weights ('clicks'), it might be worthwhile to include in the computation of similarity only the most important concepts, e.g., the top 10 concepts or the concepts having weight above a certain threshold. The exact number of concepts has to be determined in experiments.

The algorithm consists of two loops: one over the concepts in the Item-list (i.e., list of concepts in the item's profile), searching for matches in the User-list (i.e., list of concepts in the user's profile); the other loop is over the User-list, searching for matches in the Item-list. Within each loop, if there is no "perfect match" the search is for a match with the parent or grand-parent of the item. (There is no need to search for children and grandchildren, a time-consuming task, because the first loop finds matches from the other list of concepts.)

Legend:

- Score: total score of similarity b/w item and concept
- I-concept: a concept in Item-list
- U-concept: a concept in User-list
- w: weight of concept in User-list that is being matched.

Begin
Score=0
Repeat for each I-concept in Item-list:
Do case:
- *If I-concept is in User-list then Score= ++1*w /*"perfect match"/*
- *If parent of I-concept is in User-list then Score= ++ 2/3*w /*"partial match" type c: I2=U1 or I3=U2/*
- *If grandparent of I-concept is in User-list then Score= ++ 1/2*w /*"weak match" type e: I3=U1/*

End case.
Until end of Item-list.
Repeat for each U-concept in User-list:
Do case: /*no need to check again for "perfect match" between concepts of same item and user profiles/
- *If parent of U-concept) is in Item-list then Score= ++ 2/5*w /*"partial match" type b: I1=U2 or I2=U3/*
- *If grandparent of U-concept is in Item-list then Score= ++ 1/3*w /*"weak match" type d: I1=U3/*

End case.
Until end of User-list.
End.

Notes:

1) The scores for each type of match used in the algorithm are given as examples, as described above.

2) Not all user concepts must participate in the computation; as said, only the n-top concepts might be considered.

3.3. Example

The following example demonstrates the application of the filtering method using a few simulated items' profiles and a user's profile. The calculations are based on the matching scores demonstrated above.

Items' Profiles:

Item #	Ontology concepts representing the item*
Item 1	Crime → Laws
	Unrest → Civil unrest → Social conflict
Item 2	Sport → American Football
	Health → Injury
Item 3	Science → Natural science → Astronomy
Item 4	Life style and leisure
	Disaster and accident → Emergency incident

A User's Profile:

Ontology concepts in the user's profile	Number of clicks (weight)
Sport	20
Health	12
Crime → Laws → Criminal	3
Unrest	10
Lifestyle and leisure → Fishing	8

** An arrow represents parent-child relationship. The item's profile includes only the lower-level concepts.*

The application of the algorithm yields the following rank ordered list of items:

Item #	Ranking score
Item 2	0.40
Item 1	0.11
Item 4	0.06
Item 3	0.00

It can be observed that Item 2 gets the highest score because its profile includes 'American football', a child of 'Sport' in the user's profile; and 'Injury', a child of 'Health' in the user's profile – and both concepts in the user's profile have relatively high weights. Here is the exact computation of the ranking score, assuming we consider the scoring scheme in which I2=U1 $\rightarrow$ 2/3:

$$IS = \frac{\frac{2}{3}\cdot 20+\frac{2}{3}\cdot 12}{20+12+3+10+8} = 0.4$$

Item 1 gets the second highest score because of the 'close match' between its 'Laws' concept and 'Crime' in the user's profile, and also because of the 'weak match' between its 'Social conflict' concept and 'Unrest' in the user's profile. Item 1 gets a lower ranking than Item 2 because of two reasons: 1) lower scores of similarity; 2) lower weight of the matched concepts. Item 4 gets even a lower ranking because it has only one concept having any match with the user's profile: its concept 'Lifestyle and leisure' is a 'close match' with 'Fishing' in the user's profile. Item 3 gets a ranking score 0 because it has no match at all with the user's profile.

4. Evaluations of the Filtering Method

We plan to evaluate the filtering method in a controlled setting utilizing a prototype of the *ePaper* system. The main objective of the evaluations is to examine the effectiveness of the method, including the contribution of the various matching types (i.e. "perfect", "close" and "weak" matches) to performance, and to determine the optimal values for the various matching scores.

4.1. Measures of Effectiveness

Traditional measures of effectiveness of information retrieval systems usually include Precision and Recall. But these measures may not be appropriate for evaluating the quality of rank ordered items because the user might read only some of the top ranked items, while Precision and Recall are based on the total number of relevant or retrieved items, respectively. We are considering several rank accuracy measures which are more appropriate to evaluate rank-ordered results, and where the users' preferences in recommendations are non-binary. Following Herlocker et al. [2004], we are considering the following measures:

- *Rank Correlations*, such as Spearman's ρ and Kendall's Tau, which measure the extent to which two different rankings agree independent of the actual values of the variables.
- *Half-life Utility* metric, which attempts to evaluate the utility of a ranked list to the user. The utility is defined as the difference between the user's rating for an item and the "default rating" for an item.
- *NDPM Measure*, which is used to compare two different weakly-ordered ratings.

4.2. What will be Evaluated

The evaluations will include the following objectives:

1. Determination of the matching scores: The filtering method assumes different matching scores to the various possible types of matching between concepts in the user's profile and the item's profile: the highest score (1) is given to a "perfect" match, while a "close" match and a "weak" match get lower scores, considering also the direction of the hierarchical relation between the concepts (i.e., whether the user's concept is more general or more specific than the item's concept). This part of the experimental evaluations is aimed to determine the optimal scores for the different types of match.
2. Evaluation of the contribution of the various types of match between user concepts and item concepts: It is obvious that the more common concepts appear in both the user's profile and the item's profile, and the closer the user's concepts is to the item's concepts - the more relevant is the item to the

user. The question is: what is the residual contribution of the different types of match (i.e. "closeness") to the quality of the results. For example, what is the quality of results if only "perfect" matches are considered? What is the additional contribution of "close" matches? What is the additional contribution of "weak" matches? Results of these evaluations may enable us to determine if it is worthwhile to consider all types of relatedness, or perhaps only some of them are sufficient to obtain quality results.

3. Considering more than one match between related concepts in the user's and item's profiles: A user's profile may contain concepts from various levels of one branch of the hierarchy (e.g., the profile may include the concepts 'sport' and 'football'). The question is whether all concepts along the branch should be considered when compared to the item's profile, or perhaps only the concept having the highest score*weight. (Note that the score itself is determined according to the "closeness" factor, while the weight is determined according to the number of read items which included the concept).
4. Determining the number of concepts in a user's profile to consider: A user's profile may include many concepts, each having a certain weight (as explained above). Considering all concepts in the profile might be time consuming (in terms of processing time). It is likely that concepts having low weights will not contribute much to the quality of the filtering results. We will examine the contribution of low-weight concepts in order to determine a threshold for an optimal number of concepts or for concept weights. Initially, the algorithm will consider all concepts; then we will omit certain concepts (beyond a certain number or below a certain weight) and see to what degree it affects performance.

4.3. The Evaluation Plan

We plan to conduct user studies with real users (subjects), each having a content-based profile representing his interests. Some of the subjects will have similar (overlapping) profiles and some will have dissimilar profiles, in order to find out how the filtering method affects similar and dissimilar subjects.

The subjects will read News items delivered to them by the *ePaper* system and rate each item as "interesting" or "not interesting". Alternatively we are considering to use a scale bar (say from 1 to 5) to expresses the level of interest.

Some of the items read by a subject will be used for updating his/her profile (training set), while the remaining items will be used for the various tests described above (test sets). A test set of items, rank ordered by the algorithm (in any of its variations) will be compared to the subject's ratings of the items, and the measures of effectiveness (as described above) will be applied to determine the quality of the result. As described, the algorithm will vary from test to test:

1. As a result of the first set of tests (determination of the matching scores) we will adopt the best set of matching scores of similarity; these will be used in all the subsequent evaluations.
2. As a result of the second set of tests (evaluation of the contribution of the various types of match between user concepts and item concepts) we will determine the contribution of each level of proximity relatively to the performance obtained at the prior level, and determine if it is worthwhile to consider all or only parts of match types (e.g. only 'perfect' and 'close' matches).
3. As a result of the third set of tests (considering more than one match between related concepts in the user's and item's profiles) we will determine whether all concepts along a branch should be considered or only the concept having the highest score*weight. Based on that, the filtering algorithm will be adjusted.
4. As a result of the fourth set of tests (determining the number of concepts in a user's profile to consider) we will calibrate the algorithm to consider only a certain number of concepts in the user's profile, limited by a threshold number or weight.

5. Summary and Further Research

We presented a new content-based filtering method that uses ontology for representing user and item profiles, and for ranking items according to their relevancy in the electronic newspapers domain. The method is being implemented in the *ePaper* system for personalized electronic newspaper. The filtering method considers the hierarchical distance, or closeness, between concepts in the user's profile and concepts in the items' profile.

The method can be enhanced in various aspects. One possible enhancement is to assign more importance to concepts co-occurring in items read in the past by the user. An item which includes co-occurring concepts might

get a higher score than an item including the same concepts that did not co-occur in past read items. The added value of the incorporation this enhancement will be examined before being implemented in the method.

Another possible enhancement of the method is to consider penalty scores for concepts appearing in an item but not in the user's profile. This idea, which was adopted from Savia et al. [1998], means that a concept in an item's profile which does not appear in the user's profile might be given a negative (penalizing) score. The contribution of such penalty to the quality of the filter can be determined in empirical experiments.

The proposed filtering method utilizes a 3-level hierarchical ontology of News. It can, however, be generalized to other domains with their specific ontologies; and it must not be restricted to three levels. Moreover, the method can be enhanced to deal not just with a hierarchical but also with a network-based (DAG) ontology, where a concept may have many parent concepts, not only child concepts. Another possible extension to the method is to consider more types of relations between concepts, besides parent-child and grandparent-grandchild, e.g. twins of concepts. For example, a use's profile may include 'football' while an item may include 'basketball'. These extensions will be dealt with in further research.

Acknowledgment

This work was performed as part of a Deutsche-Telekom / Ben-Gurion University joint research project.

Bibliography

[Balabanovic et al., 1997] Balabanovic, M. & Shoham, Y. Fab: Content-based, collaborative recommendation. Communications of the ACM, 40(3), 66-72.

[Blair and Maron, 1985] Blair, D.C. & Maron, M. E. An evaluation of retrieval effectiveness for a full-text document retrieval system. Communications of the ACM, 28, 289-299.

[Claypool et al., 1999] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. & Sartin M. Combining content-based and collaborative filters in an online newspaper. Proc. of ACM SIGIR Workshop on Recommender Systems.

[Dai and Mobasher, 2002] Dai, H., & Mobasher, B. Using ontologies to discover domain-level web usage profiles. Proc. of the Second Semantic Web Mining Workshop at PKDD 2001, Helsinki, Finland.

[Dumais et al., 1988] Dumais, S.T., Furnas, G.W., Landauer, T.K. & Deerwester, S. Using latent semantic analysis to improve information retrieval. Proc. of CHI'88 Conf. on Human Factors in Computing, New York: ACM, 281-285.

[Guarino et al., 1999] Guarino, N., Masolo, C. & Vetere, G. OntoSeek: Content-based access to the Web. IEEE Intelligent Systems 14(3), 70-80.

[Hanani et al., 2001] Hanani, U., Shapira, B. & Shoval, P. Information filtering: overview of issues, research and systems. User Modeling and User-Adapted Interaction (UMUAI), 11(3), 203-259.

[Herlocker et al., 2004] Herlocker, J.L., Konstan, J.A., Terveen, L.G. & Riedl, J.T. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1), 5-53.

[Khan, 2000] Khan, L. Ontology-based Information Selection. Ph.D. Thesis, University of South California.

[Knappe, 2005] Knappe, R. Measures of semantic similarity and relatedness for use in ontology-based information retrieval. Ph.D. Thesis, Roskilde University, Department of Communication, Journalism and Computer Science.

[Le Meur and Steidl, 2004] Le Meur, L. & Steidl, M. NewsML 1.2 – Guidelines V1.00. Int'l Press Telecommunications Council. Retrieved: Dec. 07, 06: http://www.newsml.org/IPTC/NewsML/1.2/documentation/NewsML_1.2-doc-Guidelines_1.00.pdf

[Magnini and Strapparava, 2001] Magnini, B. & Strapparava, C. Improving user modelling with content-based techniques. Proc. of the 8th Int'l Conference on User Modeling 2001. M. In: Bauer, P., Gmytrasiewicz, J. & Vassileva, J. (Eds.): Lecture Notes in Computer Science, 2109. Springer-Verlag, London, 74-83.

[Middleton et al., 2001] Middleton, S.E., De Roure, D.C. & Shadbolt, N.R. Capturing knowledge of user preferences: ontologies in recommender systems. Proc. of 1st Int'l Conf. on Knowledge Capture, 100-107, Victoria, BC, Canada.

[Pereira and Tettamanzi, 2001] Pereira, C. C. & Tettamanzi, A. G. An ontology-based method for user model acquisition. Soft Computing in Ontologies and Semantic Web, Berlin, Springer.

[Puustjärvi and Yli-Koivisto, 2001] Puustjärvi, J. & Yli-Koivisto, J. Using metadata in electronic publishing. Project internal publication, available at http://www.soberit.hut.fi/comet/.

[Savia et al., 1998] Savia, E., Koskinen, T. & Jokela, S. Metadata based matching of documents and user profiles. Proc. of Finnish Artificial Intelligence Conference, STeP'98.

[Shoval, 2006] Shoval, P. Ontology and content-based filtering for the ePaper project. Working Paper, BGU.

Authors' Information

Peretz Shoval (Department of Information Systems Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel; e-mail: shoval@bgu.ac.il) is a Professor at the Dept. of Information Systems Engineering of Ben-Gurion University. He earned his Ph.D. in Information Systems from the University of Pittsburgh, where he specialized in expert systems for information retrieval. In 1984 he joined Ben-Gurion University, where he founded the Information Systems Program and later on founded and headed the Dept. of Information Systems Engineering. Prior to moving to academia, Shoval held professional and managerial positions in computer and software companies. Shoval's research interests include information systems analysis and design methods, data modeling, and information retrieval and filtering.

Veronica Maidel (Department of Information Systems Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel; e-mail: maidel@bgu.ac.il) received her B.Sc. from Tel-Aviv University in 2001 and is currently a graduate student at the Dept. of Information Systems Engineering of Ben-Gurion University. Her research is on content-based filtering. This paper is part of her research.

Bracha Shapira (Department of Information Systems Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel; e-mail: bshapira@bgu.ac.il) is Senior Lecturer at the Department of Information Systems Engineering of Ben-Gurion University. She holds a M.Sc. in Computer Science from the Hebrew University in Jerusalem and a Ph.D. in Information Systems from Ben-Gurion University. Her research interests include Information Retrieval and Filtering, specializing in various aspects of user profiling and personalization. In addition, she has worked on privacy preservation while browsing and on formal models of Information Retrieval systems. She is leading research projects in these domains at the Deutche-Telekom research lab at Ben-Gurion University.

LOGIC BASED PATTERN RECOGNITION - ONTOLOGY CONTENT (2)[1]

Levon Aslanyan, Vladimir Ryazanov

Abstract: Logic based Pattern Recognition extends the well known similarity models, where the distance measure is the base instrument for recognition. Initial part (1) of current publication in iTECH-06 reduces the logic based recognition models to the reduced disjunctive normal forms of partially defined Boolean functions. This step appears as a way to alternative pattern recognition instruments through combining metric and logic hypotheses and features, leading to studies of logic forms, hypotheses, hierarchies of hypotheses and effective algorithmic solutions. Current part (2) provides probabilistic conclusions on effective recognition by logic means in a model environment of binary attributes.

1. Introduction

Pattern Recognition consists in reasonable formalization (ontology) of informal relations between object's visible/measurable properties and of object classification by an automatic or a learnable procedure [1]. Similarity measure [1] is the basic instrument for many recognition formalisms but additional means are available such as logical terms discussed in part (1) of current research [2]. Huge number of recognition models follows the direct goal of increasing recognition speed and accuracy. Several models use control sets above ordinary learning sets, others use optimization and other direct forces. Besides, more alternative notions are available to describe algorithmic properties. In existing studies the role of these notions is underestimated and less attention is paid to these components. In part (1) the attention is paid to implementing the learning set through its pairs of elements rather than the elements separately. The following framework is considered: given a set of logical variables

[1] The research is supported partly by INTAS: 04-77-7173 project, http://www.intas.be