Ontological Document Filtering

Witold Abramowicz
Jacek Szymanski
University of Economics, Poznan
Al. Niepodleglosci 10, Poznan, Poland

Abstract: Recent development in business has led to new demands for information. On the other hand the spread of Internet based information services has grown to such extent that it is often impossible to find relevant information in short time. Term based information filtering systems often are not efficient enough. Therefore a need for more efficient mechanisms has arisen. This article discusses the requirements for ontological document filtering for the purposes of Information Oriented Workflow.

Introduction

Recent development in business has led to new demands for information. Information is needed for companies to act more accurately in rapidly changing environment. On the other hand the spread of Internet based information services has grown to such extent that it is often impossible to find relevant information in short time. From not having information or having little information companies have come to having too much information, although the amount of relevant information is still too low.

Currently used term based information filtering systems often are not efficient enough. The amount of information and the fact that in natural language there is a specific meaning behind each word and each phrase makes it almost impossible for such system to give accurate results. What is more, the recipients of the effects of work of such systems are not only humans but also information systems within the companies. The role of those systems is to process results received from filtering systems in order to conduct certain tasks and actions. An example of such system is microWorkflow as discussed in [ASS04].

Therefore a need for more efficient information filtering mechanisms has arisen. As it was signaled in [ASS04] mWF system bases on ontologies. A more general approach to the idea can be taken, resulting in Information Oriented Workflow (IOWf). The approach requires the information filtering systems to be able to work with use of ontologies. Thus, the base for filtering must be changed from terms to ontologies.

First part of this article presents the notion of Information Oriented Workflow as an example of possible use for ontology based filtering. The second part of the article specifies requirements for ontological filtering to be possible. The discussion ends on a presentation of two problems connected with the issue of ontological filtering: ontology translations and relevance feedback.

Information Oriented Workflow

The idea of Information Oriented Workflow bases on the “standard” perception of workflow and is supposed to be an extension to the current state-of-the-art. IOWf can be defined as the automation of a business process which occurs frequently, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules and the execution of the process can be influenced by external information, which causes alternative virtual process execution paths (alternative workflows) to be created.

The automation of a business process is defined within a Process Definition, which identifies the various process activities, procedural rules and associated control data used to manage the workflow during process enactment and which has the process information profile.

Alternative instance can be defined as the representation of a single enactment of a process, or activity within a process, including its associated data. Each alternative instance represents a separate thread of execution of the process or activity, which may be controlled independently and will have its own internal state and externally visible identity, which may be used as a handle, for example, to record or retrieve audit data relating to the individual enactment. Alternative instances do not influence reality (i.e. they do not involve applications) and are used to compare effects of alternative reactions of the original instance to external information that influences it – they run transparently for the users of IOWf System [ASS04].

Alternative instances are created each time the original instance is influenced by external information. Alternative instances are also created for previously created alternative instances in reaction to the original instance being influenced by external information thus resulting in a structure of tree.

Many individual process instances may be operational during process enactment, each associated with a specific set of data relevant to that individual process instance (or workflow "Case"). Each process instance may have multiple alternative instances created each time the instance is influenced by external information.

Information Oriented Workflows are defined, created and their execution is managed by Information Oriented Workflow Management Systems. IOWf Management System, apart from normal Workflow Management System tasks, also manages the execution of alternative workflows and allows keeping distinction and association between workflows and alternative workflows [ASS04].

One of the specifics of IOWf is information need. This notion can be defined as the need for external information (information not set at IOWf Management System level). Information needs do not need to be constant and they may be changed as a result of alternative instances efficiency evaluation.

Information needs are expressed with IOWf information profiles. IOWf information profile can be applied to processes as well as to transitions between operations [ASS04].

Requirements for ontological filtering

In order for the systems to be able to use ontologies as the base for information filtering operations there are some requirements to be met. Those requirements result from the fact that there are many languages to describe ontologies and many approaches to ontologies themselves.

Language translatability

A language L1 is translatable to the language L2 if all the elements of the language L1 can be projected using the elements and only the elements of the language L2.

$$\forall \mathrm{l} \in \mathrm{L}_1 \; \exists \, \mathrm{f(l)} \in \mathrm{L}_2$$

If the language L1 defines a notion there must be a corresponding notion or set or notions or other language construction in the language L2 in order for those two languages to be translatable.

The above definition has been placed in the context of 'family of languages' in [ES02]. [ES02] denotes several approaches which can be used:

The mapping approach

This approach is the most direct and often used. It bases on matching expressions in one language to expressions in another language. It is characterized by existence of a function mapping expressions from one language to another [ES02].

The pivot approach

The mapping approach has the drawback of requiring transformations between any language to any another. The number of necessary transformations can be reduced by creating a single pivot language all other languages are translated to.

The layered approach

This approach assumes creation of a layered architecture containing languages with increased expressiveness. With this architecture languages can be translated into languages higher in the hierarchy without loosing information.

Ontology translatability

An ontology O1 is translatable to the ontology O2 if all the relations between its elements can be projected by relations in the ontology O2.

$$\forall r \in \mathrm{O}_1 \; \exists \, \mathrm{f(r)} \in \mathrm{O}_2$$

Ontology translatability is crucial for filtering systems to work correctly when dealing with two different ontologies describing the same part of reality. In order to be able to perform reasoning basing on its base ontology filtering system must be able to convert the ontology used by processed document.

The basic translatability model where an ontology is translated into another ontology can be further expanded by adding 1 to n and n to 1 ontology translations. These models can be presented as:

$$\forall r \in \mathrm{O}_{\mathrm{s}} \; \exists\, \mathrm{f(r)} \in (\mathrm{O}_{\mathrm{D1}}, \mathrm{O}_{\mathrm{D2}}, \ldots, \mathrm{O}_{\mathrm{Dn}}) \text{ (1 to n translation)}$$

and

$$\forall r \in (\mathrm{O}_{\mathrm{S1}}, \mathrm{O}_{\mathrm{S2}}, \ldots, \mathrm{O}_{\mathrm{Sn}}) \; \exists\, \mathrm{f(r)} \in \mathrm{O}_{\mathrm{D}} \text{ (n to 1 translation)}$$

The product of such translation would also need to show relations between destination ontologies (1to n translation) or relations between source ontologies (n to 1 translation) since mere translation function might loose important relations in this case.

Ontology translations

There can be two approaches taken when dealing with translating ontologies: translation by mapping and translation by definition creation.

Translation by mapping

Translation by mapping assumes that there exists relation and element in the target ontology which can be mapped to the given relation and element in the source ontology. If the relation types in both ontologies are the same then both elements taken into consideration can be assumed to be synonyms.

Translation by definition creation

Translation by definition creation can be used whenever translation by mapping is impossible (there is no direct mapping between a relation and element in the source ontology and a relation an element in the target ontology). Translation by definition creation assumes using higher-level notions defined in the target ontology and relations to describe a lower level notion, just like definitions are created. An example of translation by definition would be to say that a *son* (lower level notion) is a *boy* (higher level notion) who *is a child of* (relation) *a mother* (specification).

Relevance measurement

In order for filtering systems to measure document relevance they must be able to compare them to user information profiles. User information profiles can be set up not only basing on pure terms but also based on ontologies. Therefore documents must also relate or be related to some ontology in order to enable comparing them to the profiles.

There are two possible approaches to this problem: natural language processing and ontological document contents mapping.

Natural language processing

There is a variety of document formats in the Internet. Most of documents can however be described by text. Therefore one could apply natural language processing to analyze the text and map it to given ontology. Such processing would require deep text analysis in order to correctly determine words' meaning and context [Ba95][BC92][Fo92].

So far natural language processing has gone a long way since it beginnings but it is said frankly that it still needs a lot of work as a technology to be able to serve for such purposes.

Ontological document contents mapping

Another approach to the problem of relating document contents to ontologies is ontological document contents mapping. This approach assumes that documents are build like markup language documents, for instance XML documents.

Words used in sentences would be the data in such a document. Behind it there would be references to base ontologies. This way each word or at least each term in the document will be linked to its base ontology. An instance sentence "Computer is a tool" could be coded in an XML document like the following schema:

<ref_to_ontology_where_computer_is_defined>**Computer**</ref_to_ontology_where_computer_is_defined>**is**<ref_to_ontology_where_tool_is_defined>**a tool**</ref_to_ontology_where_tool_is_defined>

By analyzing references filtering mechanism can match terms to proper ontologies. This approach has four major advantages. The first of them is that it enables filtering systems do their relevance operations basing on ontologies (providing that the document's base ontologies are translatable to the system's ontologies. The second advantage is that there can be word processing systems created which use provided ontologies and automatically map words to them. What is more, such systems can impose the logics from ontologies on the person writing text so that a consistent meaning of terms is maintained. This is important in organizations having numerous authors having to keep a uniform style. Thirdly, there can be no misunderstanding due to homonyms or incorrect word use. The last but not least advantage is that the level of complexity of operations required to process a document is not much higher than with processing based only on terms.

Basing on this mechanism documents can be ontology indexed. Term to ontologies references can allow creating document index vector. Document index vector can then be compared to user profiles using similarity measures. Such measures will however differ from those used in term based profiles (like Jaccard or Cosine). This results from the fact that for ontology based profiles one has to compare not only notions as terms but also relations between those notions. Therefore statistical frequency of notions appearance has to be modified by “relevance influence” of relations between them. This way the defined relations will not be lost in the comparison mechanism.

Relevance feedback

Relevance feedback mechanisms are valuable for information profile improvement. Basing on user marks for given documents those mechanisms can suggest or conduct automatically changes to user profile in order to better reflect actual user information needs.

In profiles based on terms such changes can consist on adding or removing terms or changing values of term weights or desired frequency.

For profiles based on ontologies relevance feedback mechanisms will also perform a vital role. What is more, their role can be extended because of ontology specifics.

Profile improvement

Relevance feedback mechanisms can improve ontology based information profiles. The improvement can consist on adding or removing relations between used terms and ontologies or introducing new ontologies into the profile or removing exists ones [MY94].

Ontology improvement

Relevance feedback mechanisms can also improve ontologies lying behind information profiles. With this attitude, not only filtering mechanisms can be improved but also users or corporate knowledge base extended or amended. There can be two kinds of ontology improvement: relation type change and element addition/deletion [ASS04].

Relation type change

Various ontology description languages define different relation categories between its elements. It can be so that an ontology has been created in such a way that it does not reflect the reality. It can also occur that the part of reality described by an ontology has changed and therefore the ontology is not up-to-date. Of course, the severity of such inconsistence can vary but it can cause the filtering mechanism to mark irrelevant documents as relevant and therefore provide users with low quality information. Therefore keeping the base ontologies for filtering system consistent and up-to-date is vital for the system performance.

Those inconsistencies can be a result of incorrect relation type between elements inside the ontology. One could think of an example ontology describing family relations. In such ontology it should be defined that a grandfather is the father of mother or father. If for some reason grandfather was defined as the mother of father or mother that the reasoning based on this relation would be incorrect, thus possibly resulting in incorrect relevance assessment. This example is a very improbable one but it shows the problem.

Real life ontology applications will deal with more subtle inconsistencies but such inconsistencies would be much more difficult to find out.

Relevance feedback mechanism would have to be able to assess that user marks for given document can suggest that there is a possible difference with how the user perceives the reality and how this perception has been formalized into the base ontology. Those mechanisms should suggest a proper change to the ontology to better reflect user information needs.

Element addition/deletion

Relevance feedback mechanisms can also analyze currently used ontologies in order to be able to suggest changes. As it was said before, the reality described by an ontology can change causing the ontology to be not up-to-date. Some elements of existing ontologies may be no longer used. There may be some new notions to be added to the ontologies.

The evaluation of each ontology element can be done based on information submitted by users via relevance feedback mechanisms. If documents referencing particular ontology elements are found irrelevant by the user it is possible that the element should be dropped out of the ontology. It does not mean that the ontology reflects reality incorrectly but it does mean that the ontology does not reflect the way the user perceives the reality.

It is however worth mentioning that relevance feedback mechanisms and evaluation mechanisms used in connection to relevance feedback can be very resource consuming. Such operations can be costly in terms of processor time and memory usage. Ontology based filtering will require much more sophisticated relevance feedback mechanisms and thus will consume more system resources.

There is another difficult point regarding relevance feedback in ontology based information filtering. As far as term based profiles are concerned it is not that difficult for an average user to understand how the profile works and should be created. Ontology based profiles will be more complicated compared to term based ones and reasoning based on them will be more sophisticated. In most systems profile improvement mechanisms will not be allowed to change profiles automatically, they will have to have the approval of the user. Therefore users of such systems will have to make decisions about those systems' suggestions which will require more knowledge about how ontological based information filtering works.

Summary

Ontological document filtering allows more precise and accurate document evaluation leading to submitting users with more relevant information. Ontology based relevance feedback mechanisms can be used not only to improve user information profiles but also to improve base ontologies used for those information profiles creation. Such approach allows automatic or semi-automatic actualization of ontologies reflecting changes in the real world.

This functional advantage of ontology based filtering compared to term based filtering results from the fact of extending the base for both filter building as well as relevance measurement. Since ontologies can contain not only terms but also describe relations between them, the terms can be not only a list of potentially interesting notions but they can be put and used in a deeper context.

Relations between notions can be used to add another dimension to relevance measure mechanisms. What is more relations allow ontology translations making it possible to base relevance assessments also use ontological documents content mapping.

Use of ontologies for information filtering purposes requires more sophisticated mechanisms. What is more, currently available ontology description languages have different possibilities and are not always compatible to use. Therefore the issue of language translatability arises. What is more, different organizations can have different ontologies to describe the same part of reality. Condition of ontology translatability has to be fulfilled in order to enable proper document analysis by filtering systems.

The suggested solution with ontological document contents mapping allows simplifying processing, imposing logics in ontologies on document author and reduce possible misunderstandings or misinterpretations of documents resulting from incorrect word use or understanding or from use of homonyms.

Ontology based filtering can perform a very important role in connection with Information Oriented Workflow Management Systems. The added value brought by this attitude to filtering can make IOWf achieve much better results in openness to information coming from sources external to companies using it.

References

[Ab01] Abramowicz W., Information Filters Supplying Management Information Systems, WebNet 2001, World Conference on the WWW and Internet, Orlando, Florida, October, 24-27, 2001, invited speaker

[ASS04] Abramowicz W., Stankowski F., Szymanski J, Mikro-przepływy pracy – automatyzacja małych procesów sterowania

[AS01] Abramowicz W., Stankowski F., *Instancing Workflows with Information Filtered from Internet,* A. J. Baborski, R. F. Bonner, M. L. Owoc (eds)., Knowledge Acquisition and Distributed Learning in Resolving Managerial Issues, Mälardalen University Press, 2001, 145-155, ISBN 91-88834-22-0

[Ba95] Information Filtering And Retrieval: Overview, Issues And Direction; Basis for a Panel Discussion (Moderator: Nicholas DeClaris)

[BC92] Belkin Nicolas J., Croft Bruce W. Information filtering and information retrieval: Two sides of the same coin?, 1992

[ES02] Euzenat J., Stuckenschmidt H., The ‚Family of Languages' Approach to Semantic Interoperability

[Fo92] Foltz Peter W. *Personalized Information Delivery: An Analysis of Information Filtering Methods* http://www-psych.nmsu.edu/~pfoltz/cacm/cacm.html

[Gu98] Guarino N., Formal Ontology and Information Systems. Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998. Amsterdam, IOS Press, pp. 3-15, 1998

[Gr93] Gr93er Thomas *What is an ontology* http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

[MM99] Marinilli M., Micarelli A., Sciarrone F. A Case-Based Approach to Adaptive Information Filtering for the WWW

[MY94] Masahiro Morita, Yoichi Shinoda Information filtering based on user behavior analysis and best match text retrieval (Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval pp. 272-281)

[SM83] Gerard Salton, Michael McGill Introduction to Modern Information Retrieval; 1983 McGraw-Hill Book Company

[Ta96] Takkinen Juha Information Retrieval and Information Filtering (IRIF), Spring 1996: Introduction to Course, 1996

[ZU96] Zaenen Annie, Uszkoreit Hans *Language Analysis And Understanding*