Information Filtering as a Knowledge Organization process: techniques and evaluation

Ismaïl Timimi, Stéphane Chaudiron

Information Filtering as a Knowledge Organization process: techniques and evaluation

Ismaïl Timimi, Stéphane Chaudiron
Lab. Geriico - Université Lille 3, France

Abstract

In this study, we are concerned by a field which represents an intellectual, social, and economic practice, strongly linked to a semi-automatic knowledge organization. *Informational Competitive lntelligence* is characterized by two major distinctive features: transition from the classical activity of Information Retrieval to organised lnformation Filtering, then conversion of filtered information into Knowledge to help decision making.

In the paper, we first show that information filtering systems may be considered as semi-automatic knowledge organization devices in the business intelligence context. Then, we point out how the technical dimension of the system must be arranged with the user dimension in order to approach a real relevance. Finally, we present the overview of the lnfile evaluation campaign which represents an attempt to validate our approach.

1. Introduction

Beyond the various ways of defining and explaining what is knowledge, the Knowledge Economy represents a major concern for the specialists of the domain (researchers, practitioners, economists…). This economy can not grow without paying attention to the various steps of the "knowledge chain", from automatic or human data acquisition to knowledge organization and its different uses (documentation, competitive intelligence, knowledge management...).

Knowledge Economy also face with problems of information overload at the digital age (proliferation of resources and supports, diversification of formats and structures, increase in volumetry and number of users, multilingualism requirement, and the emergence of new editorial practices…). Dealing with the consequence of overproduction means to develop and use new technologies such as clustering, push, filtering, cartography and so on with a number of components (linguistic, statistic, structural…). However, the mediation of these technologies is not without complexity and requires combination, not always obvious, between two dimensions of knowledge organization (Kolmayer, 1999):

- The technical dimension which is based on different conceptual models, various technical environment and resources…
- The user dimension which is closely linked to culture and philosophy, knowledge and know how in the field, practices and individual interests, preferences and subjectivity…

The profitability of systems depends on the compatibility and interactivity between these two dimensions. In our paper, we focus on two case studies: the case of competitive intelligence as a sub-field of knowledge economy and the case of filtering systems as mediation tools for knowledge organization.

Colloque international ISKO « Culture and Identity in Knowledge Organization », 5-8 août 2008, Montréal, Canada, Ergon, p. 367-373.

We present how filtering devices with its various by-products can be exploited in an activity of competitive intelligence and business intelligence process. Then, we discuss the two different perceptions of relevance, according to the technical approach and the user-oriented approach, in order to find out evaluation criteria which combine these two different knowledge organizations. We conclude by a presentation of the evaluation protocol of the InFile campaign devoted to information filtering systems.

2. Filtering devices in a business intelligence process

Because knowledge represent a large part of the intangible goods of each company and a way to compete more efficiently, strategic information systems and knowledge management systems are of great importance. In front of the critical proliferation of electronic information and the underlying difficulty to manage this information in a relevant way, the usual answer is to reduce drastically the volume of documents available to the end users using abstracting or filtering process (Chaudiron & Fluhr, 2001).

The functionality of filtering systems is to successfully separate relevant and non-relevant documents in an incoming stream of textual information. According to Belkin and Croft (Belkin & Croft, 1992), an information filtering system is a system designed to manage unstructured or semistructured data. We may consider that, nowadays, these systems also manage unstructured data such as pure textual documents. Information filtering systems deal primarily with textual information, involve large amounts of data incoming through permanent streams such as newswire services. Filtering is based on individual or group information profiles which assume to represent consistent and long-term information needs. From the user point of view, the filtering process is usually meant to extract relevant data from the data streams, according to the user profiles.

Information filtering systems may be used in different business environments: for example, text routing involves sending relevant incoming data to individuals or specific groups, categorization process aims at attaching one or more predefined categories to incoming documents, or anti-spamming tries to remove « junk » e-mails from the incoming e-mails. In the context of competitive intelligence, information filtering may be considered as a very specific subtask of the information management process. In this approach, the information filtering task is very similar to Selective Dissemination of Information (SDI), one of the original and usual functions assumed by documentalists and, more recently, by other information intermediaries such as technological watchers or business intelligence professionals.

As many authors mentioned it, information filtering is a key issue in the business or competitive intelligence process. In the different models of the competitive intelligence cycle, we constantly find the “information acquisition” step as a main task of the whole process. According to Boutheillier and Shearer (Boutheillier & Shearer, 2003), a specific subtask of the “information acquisition” task is to “filter [the content] in order to retain the desired information and discard unwanted information”. Filtering means examining whether the collected information address the needs, topics and requirements that were identified previously. For AFNOR (Afnor, 1998), the French official body in charge of the normalization process which provided the *de jure* standard concerning the watch services in business environment, the whole cycle of

competitive intelligence implies 8 steps among which the “information gathering and selecting” task.

While in information retrieval, systems deal with a relatively stable document set and constantly new queries, in information filtering (also known as routing or selective dissemination of information), the queries (or profiles) are fixed over the time and new documents are constantly added to the initial set. A good example of this situation is a system filtering wires coming from news agencies such as Reuters, Bloomberg or Agence France Presse (AFP).

The filtering task may be assumed by different means, included automatic tools such as filtering software but not only. More generally, filtering is a process of organizing information according various criterias. This process may be the fact of a single person (cognitive filtering) or in a cooperative way within a group or a community (social filtering). Cognitive filtering is a process that uses content of information to define the user profile. The profile contains information concerning the user’s interests and supposed information needs. The filtering technique matches coming document with the profile and the global performance of the system is evaluated through feedback from users. Information is recommended on the basis of feedback, recommendations, and cognitive profile of ‘similar’ users. In this respect, social filtering is also content-based but this model mainly uses social parameters such as a user’s education, occupation, knowledge and experience as well as preferences and habits. The system also assumes that users with matching social parameters will also share preferences and habits. This relies of the creation of user stereotypes, with sets of rules applied to each stereotype. This kind of systems usually provides ranking filtering so irrelevant items are not discarded but given a low ranking.

The process of filtering may be based on the characteristic of the document such as the words it contains (keywords which may be terms or concepts, named entities), syntactic patterns which represent events (mergers and acquisitions of companies for example) or based on a complete linguistic analysis of the document. Another way to filter documents, commonly referred to as “recommender systems” is to base the filtering process on annotations made to the documents by other users. This distinction between content based and annotation based partially meets the former distinction between cognitive and social approaches. With the development of the collaborative filtering mechanisms within specialized communities (professional or not), the question of the user model is redefined. User models are usually hand-crafted and/or refined with machine learning techniques using explicit or implicit relevance feedback.

Another approach to consider information filtering is to distinguish between text classification and text clustering. These techniques have been reported extensively in the traditional IR literature. Text classification is the classification of textual documents into predefined categories (supervised process) and text clustering groups documents into categories defined dynamically, base on their similarities (unsupervised process). In classification, categories are first determined (such as the Library of Congress Classification, the Dewey Decimal Classification or the Yahoo! categories) and the incoming information (or documents) are filtered according to a existing structured hierarchy. In clustering, categories are revealed in a bottom-up approach as result of grouping objects based on similarities. Both classification and clustering are filtering techniques

3. A relevance of filtering based on system knowledge and user knowledge

Information filtering process differs from information retrieval by several aspects but the two processes strongly agree on the problematic question of relevance concerning the results given by the systems (Belkin & Croft, 1992) (Berti-Equille, 2002).

In the case of information retrieval, the general organization consists by comparing in a single session, the query formulated in the search language of the system with the index representing the texts collection. The matching can be exact (boolean model) or optimal (vectorial or probabilistic model) with possibly a weighted answers ranking. Several tests of improvement by techniques of requests reformulation were proposed (Ben-Ali & Timimi, 1999), but this approach is still faced with problems of adequacy between the expression of the information requirement and the information presentation. Always in order to decrease the limits, another technique consists in evaluating the texts returned in this first session by the user, then reinjecting in a second session, new relevance criteria. This technique often involves a modification of the query and its progressive refinement by a process of “relevance feedback”.

In the case of a information filtering, the user formulates what is required (positive profile) and what is not required (negative profile) in a dynamic and regular information flow, using a representation of its relatively stable centers of interest on the long run. Several tests and techniques were implemented to improve the performances of filtering systems (adaptive filtering based on the progressive and iterative training, passive collaborative filtering based on the analysis of the user’s behaviors, active collaborative filtering based on user comments or analysis…). However, the relevance question still remains a big concern.

We may point out two different approaches of relevance (Denos, 1997) related to two different knowledge sources: relevance to a subject and relevance to a user.

The first "system-oriented" is based on the topic adequacy (topicality) between required information and information returned. It remains formal and mechanical and depends on the correspondence made by the system between the presentation of the request and that of the database. The second "user-oriented" is based on the decision of the user to accept or reject the information collected. It remains difficult to be identified, considered ambiguous or multifaceted to be formalised (Brouard & Nie, 2000). The user decision is mainly related to his *explicit knowledge* that is organized, in a visible way, in the form of profiles, and especially to his *tacit knowledge*, organized but in an invisible way, in his memory, his practices and his behavior.

The system efficiency not only depends on the topic exactitude of the question-answers (objective answers and modelisable organization) but also on adequacy between responses and user requirement (subjective answers and formalisable difficult process).

According to Saracevic (Saracevic, 1975), it is difficult to between these two kinds of relevance which are complementary. Green (Green, 1995) considers that the user is the real judge of relevant document but, on the other hand, it is perhaps not the best because he does not necessarily have the knowledge required to evaluate the relevance of the document. Other researchers, Schamber (Schamber, 1991) and Barry (Barry,

1994) try to determine an inventory of the useful criteria in the evaluation of the relevance by the user[1]...

In theory, the good way to find out useful criteria to evaluate systems seems to combine the system-oriented relevance criteria and the user-oriented relevance criteria. That is what we presently try to do in the InFile project, taking into account user preferences based on observations of what we call the "ground truth".

4. Main features of the InFile Evaluation Campaign

The InFile[2] evaluation campaign (INformation, FILtering, Evaluation) is a cross-language adaptive filtering evaluation campaign, sponsored by the French National Research Agency. The campaign is organized by the CEA-LIST, ELDA and the University of Lille3-GERiiCO. It has an international scope as it is a pilot track of the CLEF[3] 2008 campaigns. For those familiar with TRECs filtering tasks, the InFile campaign is similar to the TREC-11 filtering track with some characteristics (Robertson & Soboroff, 2002). InFile mainly consists of an adaptative filtering task which tries to simulate an on line crosslingual filtering process. English, French and Arabic were concerned by the process but participants could have been evaluated on mono or bilingual runs.

As a consequence of what we have previously said concerning the information filtering process in sections 2 and 3, we paid a particular attention to the context of use of filtering systems by real professional users. Even if InFile is mainly a technological oriented campaign, we constantly tried to adapt the protocol and the metrics, as close as possible, to the so-called « ground truth ». In respect with that, the global features of InFile are:

Corpora:

A newswires corpus was provided by the Agence France Presse (AFP). This is a collection of about 1,4 millions newswires (10 GB) selected from a 3 years period. Newswires are available in the three mentioned languages but are not translations from a language to another.

A set of 50 profiles was prepared covering two different categories: the first group deals with general news and events concerning national and international affairs, sports, politics… and the second one deal with scientific and technological subjects. In order to be as close as possible to the "ground truth", profiles were constructed by competitive intelligence professionals from INIST[4] (the French Institute for Scientific and Technical Information Center), ARIST Nord Pas-de-Calais[5] (Agence Régionale d'Information Stratégique et Technologique), Digiport[6] and ONERA[7] 30 of these are general profiles and 20 are scientific profiles. The practitioners constructed both the

[1] Document content and source, user's philosophy and preferences, other sources (consensus, external verification), document cost and accessibility...

[2] http://www.infile.org

[3] http://clef-campaign.org

[4] http://international.inist.fr/

[5] http://www.aristnpdc.org/

[6] http://www.digiport.org

[7] http://www.onera.fr

English and the French versions of the profiles while the Arabic version was translated by native speakers.

Relevance judgments:

The relevant set of documents was constructed in two phases, a pre-submission phase and a post-submission phase of judgements. Extensive searches using different retrieval systems were conducted at ELDA after the elaboration of the profiles. In this pre-submission phase, both the professional involved in the definition of the profiles and other assessors made relevance judgments on the outputs of the systems. This process included several feedback stages. After one round of such assessment, relevance information was used to improve the profiles and another round of assessment was made. In a post-submission phase, additional relevance judgments are planed to be made by the assessors after submission of results by the participants, on the documents taken from the pooled submissions for each profile. It will allow to identify additional relevant documents that could have been not found by the assessors at the previous stage.

Protocol and metrics:

In order to minimize a human intervention during the test, the evaluation task was performed using an automatic interrogation of participating systems with a simulated user feedback but systems were allowed to use the feedback at any time to increase performance. For each profile, systems were given a Boolean decision for each document. Due to the many possible runs, participants were also asked to fulfill a form to precise which languages and which kind of profiles they wanted to be evaluated on.

Three different metrics have been retained:

- *Progression measure* (or evolutivity) which measures the ability of a system to improve itself from the relevance feedbacks;
- *Originality measure* which measures the fact that a system is the only one to retrieve some relevant documents;
- *Anticipation measure* which measures the ability of the systems to retrieve the first relevant document; this measure is very closed to real conditions of use when it is important to extract “low signals” from an incoming flow of information.

These metrics try to take into account the user information behavior during the relevance judgment phase. The metrics have been elaborated after discussions with CI practitioners. They surely don’t fit exactly with the real conditions of use but hey ca be considered as a first attempt to match with these conditions.

5. Conclusion

At this time, the real test of the InFile campaign didn’t start yet, so we are not able to present the results of the comparative evaluation of the participants but the first goal has been achieved. This goal was to define an evaluation protocol paying attention to a real context of use.

Information filtering systems can be considered as a case study to demonstrate how it’s possible to deal with a user evaluation referring to cognitive and psychosocial

influences and a technical-functional assessment in a unified approach, in order to evaluate systems.

References

Afnor, 1998. Norme XP X 50-053. Prestations de veille et prestations de mise en place d'un système de veille.

Barry, C.L. 1994. User-defined relevance criteria : an exploratory study. *J. of the Americain Society for Information Science*, 45(3), p. 149-159.

Belkin, N. & B. Croft. 1992. Information filtering and information retrieval: two sides of the same coin, In *Communications of the ACM*, december 1992, vol. 35, n°12, pp. 29-38.

Ben Ali, S. & I. Timimi. 1999. De la Paraphrase à la Recherche d'Information : Système 3AD. *Colloque International en Sciences de l'Information* (*CISI'99*), Thème : les bibliothèques à l'ère des réseaux d'information, Tunis, texte n° 3, 16 p.

Berti-Equille, L. 2002, Annotation et recommandation collaboratives de documents selon leur qualité. *Revue des sciences et technologies de l'information*, série ISI-NIS, vol.7, n°1-2/2002, p. 125-155.

Boutheillier, F. & K. Shearer. 2003. *Assessing Competitive Intelligence Software : A Guide to Evaluating CI Technology*, Medford, Information Today Inc..

Brouard, C. & J-Y. Nie. 2000. The system RELIEFS : a new approach for information filtering. *Proceedings of the Text Retrieval Conference (TREC-9)*, Gaithersburg, p. 513-517.

Chaudiron, S. & C. Fluhr (sous la dir. de). 2001. Filtrage et résumé automatique de l'information sur les réseaux - *Actes du 3ème Colloque du Chapitre français de l'ISKO*, Nanterre, Université de Paris X, 283 pages.

Denos, N. 1997. Modélisation de la pertinence en RI : modèle conceptuel, formalisation et application. Thèse de Doctorat, UJF, Grenoble 1.

Green R., 1995. Topical relevance relationships. I. Why topic matching fails, *JASIS'95*, 46(9), pp. 646-653.

Kolmayer, E. 1999. Knowledge organization and expertise among the users , In *Proceedings of International Society of knowledge Organisation (Isko'99)*, Lille-France, pp. 355-366.

Robertson, S. & I. Soboroff. 2002. The TREC 2002 Filtering Track Report. In Proceedings of The Eleventh Text Retrieval Conference (TREC 2002), NIST Special Publication : 500-251, http://trec.nist.gov/pubs/trec11/t11_proceedings.html.

Saracevic, T. 1975. Relevance: A review of the literature and a framework for thinking on the notion in information science, *JASIS*, pp. 321-343.

Schamber, L. 1991. users'criteria for evaluation in a multimedia environnement, *Proc. of the Americain Society for Information Science* (*ASIS'91*), p. 126-133, 1991. Relevance and information behavior, *Arist*, 29:3-48, 1993.