

Quality, Relevance and Importance in Information Retrieval with Fuzzy Semantic Networks

Roy Lachica¹, Dino Karabeg² and Sasha Rudan³

¹ Bouvet ASA, Norway, roy.lachica@bouvet.no

² Department of Informatics, University of Oslo, Norway, dino.karabeg@ifi.uio.no

³ HeadWare Solutions, Serbia, sasa.rudan@gmail.com

Abstract. We propose a framework for ranking information based on quality, relevance and importance, and argue that a socio-semantic contextual approach that extends topicality can lead to increased value of information retrieval systems. We use Topic Maps to implement our framework, and discuss procedures for calculating the resource ranking. A fuzzy neural network approach is envisioned to complement the process of manual metadata creation.

Keywords: Topic Maps, quality, relevance, importance, semantic search, ontology, resource ranking, information retrieval, contextual search, scoping, tagging, knowledge based systems.

1 Introduction

The Web has enabled an explosive growth of information sharing, but it has also escalated the problem of information overload. The challenge that is now before us is to identify valuable information as judged by the individual user and present the end user with the right information at the right time and place. Organising information by such technologies as Topic Maps answers this challenge only partially, because among the provided topics, associations and resources, some will always be more valuable than others and have different value for different people. In this article we propose a framework for ranking information based on three criteria—quality, relevance and importance—and offer a compound measure called QRI as an extension that can improve the value of information retrieval (IR) systems.

The main objective of IR is the retrieval of relevant information [1]. IR thus becomes a particularly important area for socio-semantic systems, where perceived irrelevant information has been singled out as a key obstacle to metadata creation [2]. Another related problem is the creation of ontologies which is generally perceived as being time consuming and difficult [3]. There is also the reluctance among both users and institutions to create metadata [2]. We describe how mimicking neural networks to IR can solve these problems.

2 Defining Quality, Relevance and Importance

Our framework refines the conventional view in IR where relevance is the deciding criterion. We employ two additional criteria—quality and importance. In what follows we first survey the ways all three concepts have been treated in literature, and then define them as they are used within the QRI model.

2.1 Quality, Relevance and Importance in the Literature

Based on an analysis carried out by Knight and Burn [4], we identify reliability, availability and relevancy as the main dimensions of quality. According to this study, the quality of information is a compound criterion reflecting a number of specific characteristics.

Table 1. Categories of Information Quality

Reliability	Availability	Relevancy
Accuracy, Concise, Objectivity, Believability, Reputation, Understandability	Security, Accessibility, Navigation, Consistency	Useful, Efficiency, Timeliness, Value-Added, Usability, Amount, Completeness, (Concise)

The notion of relevance is often debated. This concept is both complex and multidimensional. However, in the field of information science, a consensus on the meaning of ‘relevance’ seems to be emerging [1]. Relevance is generally divided into two main categories: topical relevance and user-centred relevance [5]. Topical relevance is objective and mainly concerned with terminology. Topical relevance can be judged by subject area experts. User-centered relevance on the other hand is subjective to the user. Saracevic [6] defines a stratified model of relevance in IR. Relevance occurs on several connected levels. The lower levels concern the interaction with the information system while the upper levels define the user interactions. The upper level consists of: cognitive, affective, situational and contextual aspects. Situational relevance or utility is the relation between the situation, task, or problem at hand, and the resource. Affective or motivational relevance is relation between the intents, goals, and motivations of a user, and a resource. Cosijn and Ingwersen [7] define sociocognitive relevance as the relation between the resource and the situation, task or problem at hand, as perceived in a sociocultural context. At the top of Saracevic’s model is context, which is general and long term. It includes organizational, institutional, community, cultural and historical contexts. Dey [8] uses the term context for any information that can be used to characterize the situation of a user. We use the term context to refer to all the factors that determine what is relevant to a user or group.

Importance as criterion for evaluating information has received little attention in literature. Laudan [9] points at the lack of a viable framework for evaluating this concept, which in part belongs to the realms of philosophy and ethics.

2.2 Quality, Relevance and Importance in the QRI Model

As the above brief analysis shows, quality, relevance and importance have been defined in the literature in a variety of ways. A consequence of this is that the distinctions between those three concepts remain unclear.

Aiming to create a clear-cut set of criteria by which information can be evaluated and ranked by a given user in a given situation, we make the following definition for use in our Topic Maps based framework:

Quality. Quality reflects the intrinsic value of an information resource as judged by an individual. Information that is unreliable or impossible to understand is valueless, even if it may otherwise be highly relevant or important.

Relevance. Relevance is the validity of a relation between two subjects as judged by an individual in a given context. Different persons with different backgrounds might have different opinions about the appropriateness of relations between concepts.

Importance. Importance reflects the degree to which a relation between a user and a subject is valid in a given context. The perceived importance of a subject changes over time as the background and setting of the individual change.

3 Related Research

Research in the crossing point between socio-semantics and contextual information retrieval is scarce. Cantador and Castells [10] propose a multi-layered approach for social applications. Their approach compares user profiles in relation to semantic topics in order to find similarities among users.

Research on ontology based contextual information retrieval is on the other hand more widespread. Context aware relevance ranking can be found in [11], [12]. Stojanovic [13] presents a novel approach for determining relevance in ontology-based search. Siberski [14] discuss why preferences are needed in search and presents a model for use with RDF. Ontology-Based personalisation in IR has been researched by Cantador et al. [15]. Castells et al. [16] also propose the extension of an ontology-based retrieval system with semantic-based personalization techniques. Jrad et al. [17] describe an architecture that provides personalization facilities based on a contextual user model.

4 Assigning Quality, Relevance and Importance

Our model is intended for information retrieval in collaborative knowledge-based systems. Dedicated users create a shared semantic network. All subjects within the system can be used to tag resources. A central feature of the system is listing relevant topics from pages representing different subjects. Ranking of these lists as well as search results lists are seen as the main motivation of the QRI model.

We envision two ways of assigning the value of information w.r.t. each of the criteria: *manual* (by direct input or evaluation) and *automatic* (as a side effect of normal access and use).

We now turn to the central task of how users of the system add the data needed later for resource ranking. The system, for which this framework is intended, should allow users to browse subjects, users and resources. During browsing of the knowledge base the user can assign quality, relevance and importance.

4.1 Manually Assigning QRI

Users can choose to give ratings from 0 to 10 on each of the single QRI criterions. Giving a low rating will hide the topic or association for the user.

Quality. All users can rate the quality of resources. Users will have different rating influence determined by other users through cumulative popular vote. The influence of a user should be a reflection of his trustworthiness, authority, contribution and knowledge level.

Relevance. The user can rate the appropriateness of any subject to subject association. This can be understood as; if the user thinks the association makes any sense. Because relevance is context dependant we add a set of Topic Maps constructs to let the users express their context. Dey and Abowd [18] identify 4 context types:

Location. The location can easily be captured in mobile solutions, but it can also be set in stationary applications by letting the user have a set of popular locations such as 'at work' and 'at home'. The user should have the ability to add locations that are relevant to him in any way. The IP address of the logged-in user might be used to find the location.

Identity. The social and cultural background of the user can be estimated by his or her group membership. The system must therefore support sociability and group management. Identity can also be added through user profile properties like age, education, job, income, etc. although this would have some privacy concerns. The user's knowledge is also part of his or her identity. This can be supported by having such association types as 'has knowledge about' or 'is expert in'. Pomerol and Brézillon [18] provide an explanation of the relationships between knowledge and context.

Activity. The user cognitive state describes the current situation and mindset of the user. In order to support user cognitive state the system should support various activity related topics such as events, tasks, and projects.

Time. Contextual objects have a start and end date property. Part of the user profile is a local time zone property.

Importance. Users can assign importance to any topic by giving it a rating between 1 and 10. A high rating would imply that the topic is very important to him. Topics that are important to a user can be listed on the user profile page, on the start page after logging in or similar.

Importance may be set bottom-up or top-down. Bottom-up importance is added by end users. Top-down importance is added by moderators, managers or system owners who want to bias the resource ranking. In some applications it could be desirable to define important information without context. For example a panel of experts might stress awareness to climate change for all members of the system.

4.2 Automatically Assigning QRI

Automatically created QRI may be seen as suggestions made by the system. Automatically created associations receive a relevance weight of one tenth of manually created ones.

Quality. Highly ranked users authoring resources will automatically give a high quality ranking.

Relevance. Simultaneous browsing of two different topics by the same user within a short time span will create a low weight relation. Following an association from one topic to the next will increase relevance.

Importance. Whenever a user browses or use a subject it is marked as important to the user but with a low weight.

5 Topic Map Implementation

There are four master topic types in the ontology: tags, social-items, context-items and resource proxies. Tags are any free subject created by a user. A social item is either a user or a group. Instances of the tag and context item topic type are used as a category or label for tagging resources.

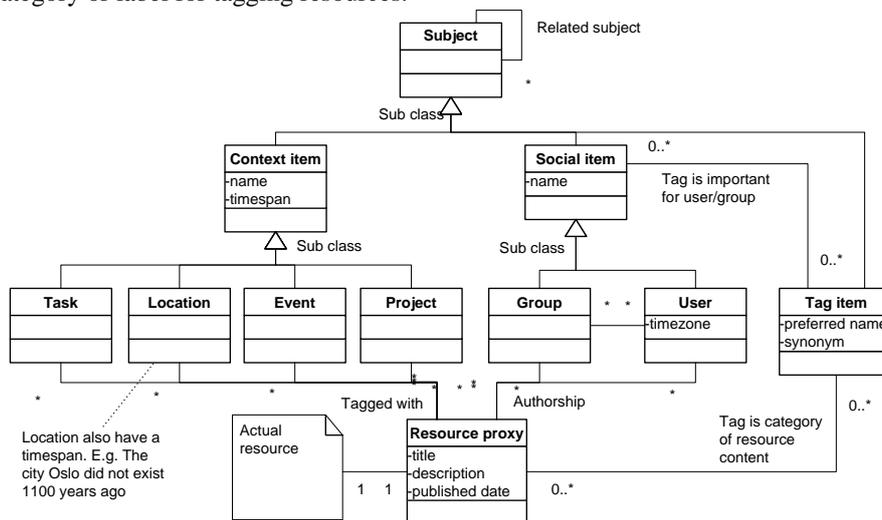


Fig. 1. Basic ontology overview.

Topic Maps provide an intuitive model for expressing topical relevance. In order to support user-centred relevance we have added the topic types; task, location, event and project along with time span and time zone occurrence types. TMCL is used to constrain allowed associations between topics.

5.1 Listing of topics, associations and occurrence types

Table 2. Topic and occurrence types.

Topic type	Occurrence types
Event	Time span
Task	Time span
Project	Time span
Location	Time span, Alias
Tag	Synonyms
Person	Local time zone, Time span
Group	
Resource proxy	

Table 3. Automatically created association types.

Topic type A	Association type	Topic type B
User	Has browsed	[Topic]
User	Has used (tagged, commented, edited)	[Topic]
User	Is browsing from (real world)	Location
User	Has created	[Topic]
User	Has browsed	[Topic]
User	Has communicated with	User
[Topic]	Single user concurrent browsing	[Topic]

Table 4. Manually created association types.

Topic type A	Association type	Topic type B
User	Is to perform *	Task
User	Is friend of	User
User	Is from *, Is currently in/at *, Has been to, Is born in	Location
User	Has authored	[Resource]
User	Has recommended resource	[Resource]
User	Has voted topic as important	[Topic]
Group	Important (favourite)	[Topic]
User	Is to attend *	Event
User	Is member of *	Group Project
User Group	Has knowledge about	[Subject]
[Subject]	Sub-class of, Type of, Is part of	[Subject]
[Subject]	Involves	[Subject]
[Subject]	Is category of (tagging)	[Resource]

[Subject] represents a social-, context- or tag item. [Resource] represents a resource proxy topic which points to an information resource through its subject locator. [Topic] represents any topic what so ever including resource proxies. Transient contextual items (*) switch their association type automatically. If a user create the relation 'User' – 'is to carry out' – 'task', the association will change to 'has carried out' when the current local time has passed the time span occurrence value of the item. When choosing to set a topic as important, an 'important for'-association between the user and the topic is created. Top down importance is created by an expert panel or similar by creating the same relation between a topic and a group.

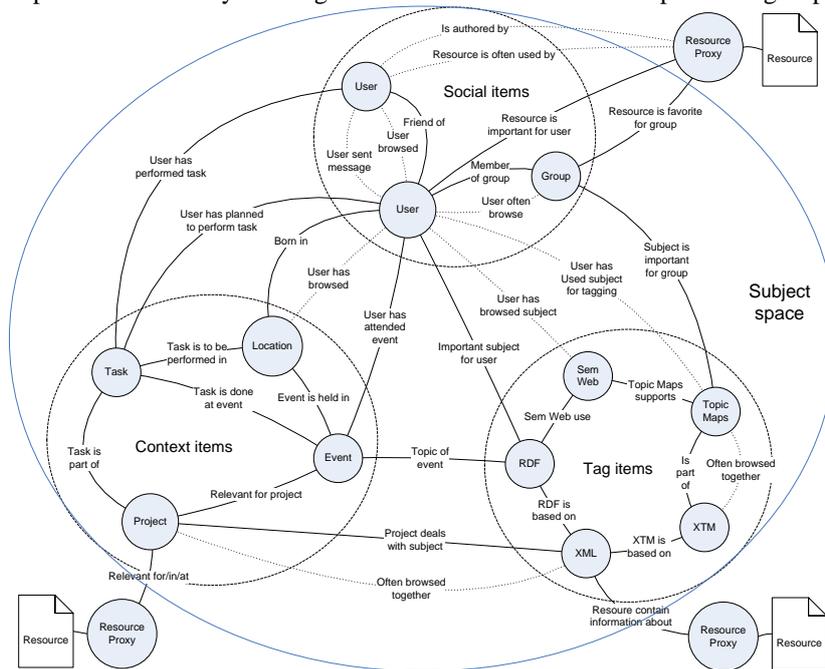


Fig. 2. Sample socio-semantic contextual network. Dotted lines show associations created automatically by the system.

5.2 Mimicking Fuzzy Neural Networks

Knowledge is fuzzy by nature [20]. We use Topic Maps to represent evolving knowledge which is lexically imprecise and/or uncertain. Topic Maps provide a good platform for evolving a knowledge structure similar to that of Collins and Quillian's Semantic Network Model [21]. Central to our neural network approach is Hebbian theory [22], which describes how associations are strengthened with use and weakened when not used. If a user clicks on a tag and he does not find it interesting, it is likely that he will not click on it again and the trail will fade away over time, thus reducing relevance. While most implementations of ontology-based IR rely on bivalent formal engineered ontologies, we utilize a promiscuous approach where users

can create semantics ad-hoc. Multiple and overlapping pathways may be created without time consuming validation or having to adhere to formal schemes that often require high cognitive load on the part of the user. The system learns what is relevant by tracking user interactions and by letting users change the network and its relevance weights. The neural network approach is also used because of the dynamic and complex nature of the user context. Context differs drastically because of surroundings, circumstance, settings, changing goals, the nuances of local and wide global influence. This makes it difficult to have up-to-date information about the context of a user [23]. Our model seeks to solve this problem by automatically evolving a context through the associations growing out from the user topic.

5.3 QRI Implementation

QRI data are kept outside of the topic map data model since it would otherwise demand extreme processing power to process the required context related queries. Also the Topic Maps data model does only allow an association to be scoped by a single topic. Our context tables described below enables a more nuanced scoping by allowing an unlimited number of weighted topics.

Quality. Quality is stored as a rating from 1 to 10 per user giving the rating on resource proxies.

Relevance. Relevance is stored in context tables. These tables will be created for associations if a user decides to rate an association. If for example two parallel associations have been created between the two same topics, the system can decide what is the most relevant for a user based on his context. The context table is populated by retrieving information about the user location, identity, activity, time and knowledge and inserting it into an array. For example a user may be related to several locations through the promiscuous semantic network at the time of defining an important item. The PSI of each location found by CSA (see section 6.2) is added to the location entry along with the semantic distance.

Importance. When an ‘important for’-association is created a context table will be added. This context table will describe in what context the subject is important for the user. This data can then be used for recommending subjects for other users sharing a similar context.

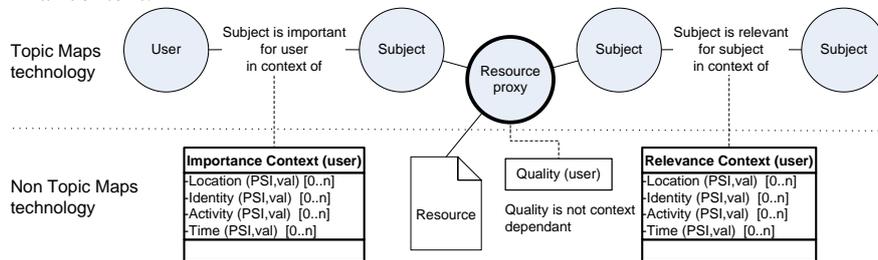


Fig. 3. Conceptual overview of hybrid Topic Map QRI implementation.

The main benefit of this approach is that the current context is captured when creating semantics. This context is then used to give the user information that makes more sense. We envision that this model can also be used to collaboratively evolve ontologies in a bottom up approach. The QRI data can be used in a filtering process to output a consensus topic map.

6 Resource Ranking Calculation

We first describe our general model for resource ranking, and then discuss three scenarios which all have in common the calculation of contextual dependant semantic distances and the use of the quality scores on resource proxy topics. We conclude this section by describing concrete implementations within the *fuzzzy.com* online socio-semantic bookmarking service.

6.1 The Basic Model

Resource ranking is calculated using semantic distance by traversing associations in the topic map. The total semantic distance is measured from the user topic. 'Important for'-associations act as entry points into the semantic network along side other relations through for example contextual topics.

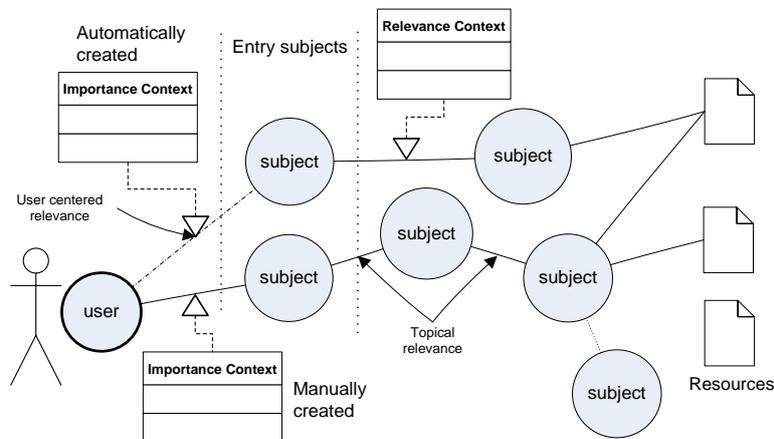


Fig. 4. Simple conceptual overview of semantic distance calculation.

The end result of the ranking will be based on the semantic distance and the quality rating of the resource. Users will have the ability to tune the influence of relevance and quality in the IR process.

Different association types have different weights. When travelling up a ‘class sub-class’ association (more abstract) the weight is decreased more than for other type of associations.

6.2 Ranking in the Context of a Specific Topic

In this scenario, ranking is calculated by following all outward paths from a start-up topic. For each hop, relevance weights are decreased by a configurable factor. All resources above a certain threshold value will be ranked as relevant. The Constrained Spreading Activation (CSA) technique [24] is used for this purpose. A second pass will increase ranking of resources that are related with the user by using the same method of outwards traversal. For contextual topics, relevance weights are automatically adjusted. The ‘Attends’-association shown in figure 3 will have its weight increased when the time of the event is near. Resources found in this process will be ranked by summarizing the relevance score and the quality score assigned to the resource.

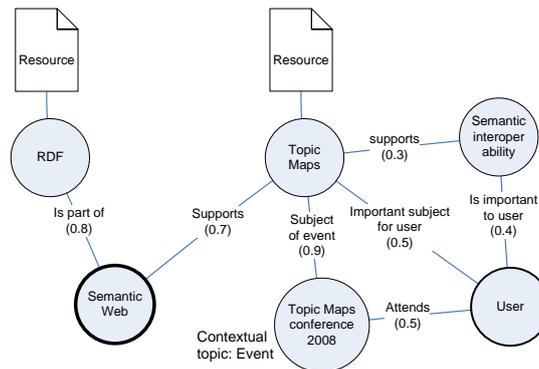


Fig. 5. The starting subject ‘Semantic Web’ has a higher association weight with the ‘RDF’ node in comparison with ‘Topic Maps’. ‘Topic Maps’ will be ranked as more relevant because it is supported by multiple pathways and it is closer in distance to the user making the request.

6.3 Ranking in Keyword Search

Each keyword in the query is matched against topics of the topic map. A syntactic term set enlargement [25] is used to retrieve matching topics by searching preferred names, aliases and using automatic singular/plural nouns. A semantic term set enlargement is performed next using the same spreading activation method as described in the previous section. If a search is performed from a particular subject page, that subject may also be used as an additional start node.

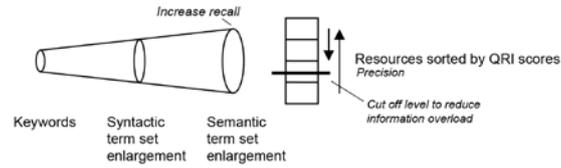


Fig. 6. The keyword search process.

For additional hits a second search may be performed using a keyword match in the resource proxy name and description occurrences.

We will now have a list of subjects that will be used to retrieve relevant resources. The ranking of resources is here calculated using a shortest path algorithm for undirected weighted graphs. The Bellman-Ford algorithm [26] may be used for this. When there are few hits a third pass should be used for retrieving resources containing the subjects using other indexing search methods.

6.4 Ranking in a Push Scenario

In a push scenario the user is not requesting information. An example of this could be an automated e-mail digest service. The system must use the available context to find what is relevant for the receiver, often referred to as Best Bets systems [27]. In the two above scenarios we can assume the user has already articulated his information needs by browsing or by query formulation. In this scenario, ranking is based on QRI, novelty and user history. The user should only receive lists of new items that he has not already viewed and which are of high importance.

6.5 Feedback loop

Before the topic map is densely populated, ranking in early stages of the system will be inefficient since the required paths between the user topic and the actual relevant resource proxy may not yet exist.

As users visit, use or rank resources, associations between the user and resource proxies are created. Again the Hebbian effect will strengthen relevant associations and less relevant will die out. The relation between the user and the resource will leave a semantic path which will allow other users to find the resource if the user share a similar context.

A timer service or similar mechanism will remove irrelevant information. For each time interval all association weights below a certain level will decrease. The time interval is configurable and should depend on the association/topic ratio.

6.6 Partial Implementation on fuzzyzy.com

Many of the ideas in this paper have evolved from the issues uncovered in the online bookmarking service fuzzyzy.com. The current version of fuzzyzy supports relevance ranking by letting users vote on associations between tags. Users can also define favourite tags, users and resources. This functionality let users directly set items as important to him, but without any context. A resource can also be voted on with a positive or negative vote to indicate quality. Fuzzyzy has a built in simple contextual semantic search feature. Upon a keyword search, keywords will be matched against all tag names in the system. All tags that have been created, used, or have been set as a favourite by the user are weighted higher.



Fig. 7. Voting on associations in fuzzyzy.

Users are able to view the relevance of associated topics as a sorted list and can move associations up or down through voting. Related items below a lower threshold are hidden. Users can create any association they like and it is up to the community to vote for or against the association.

The ideas presented in this paper will gradually be implemented on fuzzyzy.com. Tuning the QRI resource ranking is, among many other areas, a natural continuation of this project along with measuring and benchmarking precision and recall.

7 Concluding Remarks

In this paper we have shown a model for introducing quality, relevance and importance (QRI) in IR with Topic Maps. The model is designed for use in social collaborative systems where concepts such as persons, events, tasks, projects etc. are central. We hypothesize that our neural network approach to IR has the advantage of being intuitive for end-users, as associations can explicitly be shown in the user interface in comparison to other systems where the user does not know why things are listed as relevant. The burden on users to create the underlying semantic network is reduced with a neural network approach where associations are automatically created and evolved both manually and automatically.

Our model introduces a new layer on top of Topic Maps for weighted associations and for advanced contextual scoping which is intended to better support user context. All these measures together aim to provide the end users with the right information at the right time and place.

References

1. Borlund, P.: The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10), (Aug. 2003) 913-925.
2. Lachica, R., Karabeg, D.: Towards holistic knowledge creation and interchange Part I: Socio-semantic collaborative tagging. *Proc. Third International Conference on Topic Maps Research and Application, Leipzig. Lecture Notes in Artificial Intelligence, Springer: Berlin (2007)*
3. Hyvönen, E., Saarela, S., Viljanen, K.: Application of ontology based techniques to view-based semantic search and browsing. In *Proceedings of the First European Semantic Web Symposium, May 10-12, Heraklion, Greece, (2004). Springer Verlag, Berlin.*
4. Knight, S.A., Burn, J.M.: Developing a Framework for Assessing Information Quality on the World Wide Web. *Informing Science Journal*, Vol. 8, (2005) pp. 159-172
5. Kagolovsky Y, Mohr JR.: A new approach to the concept of relevance in information retrieval (IR). In: Patel V, Rogers R and Haux R (editors). *Proceedings of the 10th World Congress on Medical Informatics (Medinfo 2001). Amsterdam, The Netherlands: IOS Press, 2001 Sep;10(Pt 1):348-52*
6. Saracevic, T.: Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(3), (2007) 1915-1933.
7. Cosijn, E., Ingwersen, P.: Dimensions of relevance. *Information Processing and Management*, 36(4), (2000) 533–550.90.
8. Dey, A.K.: *Understanding and Using Context, Personal and Ubiquitous Computing*, vol. 5, no. 1, 2001, pp. 4-7.
9. Laudan, L.: *Progress and its Problems (Berkeley, Los Angeles, London: University of California Press, (1971).*
10. Cantador, I., Castells, P.: Extracting Multilayered Semantic Communities of Interest from Ontology-based User Profiles: Application to Group Modelling and Hybrid Recommendations. *Computers in Human Behavior, special issue on Advances of Knowledge Management and the Semantic Web for Social Networks. Elsevier. In press. (2008)*
11. Bénédicte Le Grand, Marie-Aude Afaure and Michel Soto. Semantic and Conceptual Context-Aware Information Retrieval. In the *IEEE/ACM International Conference on Signal-Image Technology & Internet-Based Systems (SITIS'2006), Pages 322-332, Hammamet, Tunisie, 17-21 December 2006*
12. Aleman-Meza, B., Halaschek, C., Arpinar, I. B., Sheth, A.: Context-Aware Semantic Association Ranking. Paper presented at the *First International Workshop on Semantic Web and Databases, Berlin, Germany. (2003)*
13. Stojanovic, N.: An approach for defining relevance in the ontology-based information retrieval. In: *Proceedings of the International Conference on Web Intelligence (WI), Compiègne, France (2005) 359–365*
14. Siberski, W., Pan, J.Z., Thaden, U.: Querying the semantic web with preferences. In: *Proceedings of the 5th International Semantic Web Conference (ISWC), Athens, GA, USA (2006) 612–624*

15. Cantador, I., Fernández, M., Vallet, D., Castells, P., Picault, J., Ribière, M.: A Multi-Purpose Ontology-Based Approach for Personalised Content Filtering and Retrieval. *Advances in Semantic Media Adaptation and Personalization*. Springer-Verlag, Studies in Computational Intelligence, vol. 93, pp. 25-51. (2008)
16. Castells, P., Fernández, M., Vallet, D.: An Adaptation of the Vector-Space Model for Ontology-based Information Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 19 (2) (2007), pp. 261-272
17. Jrad, Z., Aufaure, M.-A., Hadjouni, M.: A Contextual user model for Web personalization, in: *Personalized Access to Web Information (PAWI'2007)*, Nancy, France, 3-7 December 2007, 12 p
18. Dey, A., Abowd, G.: Towards a Better Understanding of Context and Context-Awareness, Workshop on the what, who, where, when and how of context-awareness at CHI 2000, April 2000.
19. Pomerol, J., Brézillon, P.: About some relationship between Knowledge and Context. Submitted to the 3rd International Conference on Modeling and Using Context (CONTEXT-01). Series Lectures in Computer Science, Springer Verlag. (2001)
20. Zadeh, L.A.: A theory of commonsense knowledge. In H.J. Skala et al., editor, *Aspects of Vagueness*, pages 257–295. Reidel, Dordrecht, 1984.
21. Collins, A.M., Quillian, M.R.: Facilitating retrieval from semantic memory: The effect of repeating part of an inference. In A.F. Sanders (Ed.), *Acta Psychologica 33 Attention and Performance III* (pp. 304-314). (1970) Amsterdam: North-Holland Publ.
22. Hebb, D.O.: *The organization of behavior*, New York: Wiley (1949)
23. Greenberg, S.: Context as a dynamic construct. *Human-Computer Interaction*, 16, (2001), 257-268.
24. Crestani, F., Lee, P.L.: Searching the web by constrained spreading activation. *Information Processing & Management*, 36(4), 2000, 585-605.
25. Kracker, M.: A Fuzzy Concept Network Model and its Applications. In: *Proceedings of the FUZZ-IEEE '92*, San Diego. pp. 760-768. (1992)
26. Bellman, R.: On a Routing Problem, in *Quarterly of Applied Mathematics*, 16(1), pp.87-90, (1958)
27. Attardi, G., Esuli, A., Simi, M.: Best bets: thousands of queries in search of a client. In *Proceedings of the 13th international World Wide Web Conference on Alternate Track Papers & Posters* (New York, NY, USA, May 19 - 21, 2004). WWW Alt. '04. ACM, New York, NY, 422-423.