

Context-aware Automatic Query Refinement Using Indian-Logic Based Ontology

S. M. BaalaMithra, S. M. SominMithraa

Department of Computer Science and Engineering, Anna University - Chennai, India

baalamithra@yahoo.com

sms.mithraa@gmail.com

Abstract - This system proposes Indian-logic ontology based Context-aware Query Refinement model to support context-sensitive semantic search in keyword based search engine. This is by formulating effective query using Indian logic based Ontology for Context identification to overcome ambiguous query terms and increase the relevancy of documents retrieved. Here we propose Indian-Logic based Context Ontology for Context representation of ambiguous keywords along with Domain Ontology for complete domain knowledge. This system effectively uses these ontologies to identify the context and iteratively refine the query. This system visualizes results based on ontology to easily obtain an overall picture of the types of results obtained and also gives indication of relationships and subcategories within the refined results. The experiment result shows that our proposed method shows better precision compared over normal search engine results.

Keywords: query refinement, query expansion, web-search, ontology, Indian-logic ontology, context.

1. Introduction

Web search enables a user to navigate through the web information content incrementally and interactively. But, the effectiveness of queries cannot be guaranteed and it varies from user to user, though the information need is common between users. So most of the users get irrelevant results or the effort for reaching needed information is high. In other words, users often do not know to combine the right words, which express their information need. Due to the lack of domain knowledge, users tend to post very short queries which do not express their information need clearly, which reduces the precision and re-call of the search. Query refinement comes in handy at those situations.

By refining a query, the web search process moves towards the goal of information need. Various methods of query refinement approaches exist in literature, but they lack semantic influence as they don't insist on user's context of search. Hence this novel system is proposed to identify the user's context of search first and iteratively

improve the relevance of web-search results using context ontologies. Our proposed approach deals ambiguous keywords by both statistical and semantic approach using Indian-logic ontology. This method statistically identifies the keywords that suggest the possible context of the user clicked page and augments it with a pre-designed Context-ontology to ensure the context. Once the context is identified it fixes the domain of the user search, and using the concerned domain-ontology the user-query is refined. Then the search result obtained using refined query is re-ranked with the matching concepts in the identified domain. The re-ranked results are visualized in a way with all the result page links pointing their context and surrounded by the main concepts identified in them. This is achieved using the domain ontology. Along with the visualized result the other possible context and possible query suggestions are also given to the user. This helps the user to manually choose the context and keyword suggestion, if he finds it deviating from his intention or wanted intentionally to change the context.

After this introduction, in the next section we give a discussion of related work, followed by technical presentation of our method. Then, we present some experimental results to prove the effectiveness of our method; the last section includes conclusions and future work.

2. Related work

Yueqin Hang et al., found ambiguity of query terms have been long-standing problem in information retrieval field, which seriously affects the effectiveness of retrieved results. A local analysis and global analysis is performed on the query and the page content to identify the context statistically [11]. This is intended for a single document mostly reflecting a concept in single context.

Query Refinement is an essential information retrieval tool that interactively recommends new terms related to a particular query. Web being unstructured and semantically heterogeneous, keyword based queries are likely to miss important results, so refining and query expansion plays an important role today [3].

A prototype of an intelligent retrieval system is provided and the running process, reason and query of the system are discussed in [8]. Ontologies are introduced to describe the semantic information of knowledge bases in order to meet requirement of utmost sharing and reuse of the domain knowledge.

Ontology based Semiautomatic Query Refinement model for the exploitation of domain ontology's knowledge bases which formulates effective query using refinement based on Ontology for increasing relevancy of the retrieval documents. [9].

[6] proposed Gautama, a tool for editing the ontology based on Nyaya logics. NORM is the Nyaya based Ontology Reference Model, which defines the standards for constructing ontology, based on the recommendations of the epistemology definitions of Nyaya-Vaisheshika school of Indian philosophy. NORM is organized as two-layer ontology, where the upper layer represents the abstract fundamental knowledge and the lower layer represents the domain knowledge

Another way of improving search result is by means of personalized-search using ontology. An ontology model for personalization is built by considering user information, user interests, preferences and his other Internet profiles [7]. This approach need to collect and preserve different information and predicting quick user interest change is difficult in this approach.

Gauch at al. [10] proposed a system that adapts information navigation based on a user profile structured as a weighted concept hierarchy. The user may create own concept hierarchy and use it for browsing web sites. A user model can also be built using an ontology schema. Paper [4] showed improvement in this methodology by coupling user interest along with concept hierarchy including low-level ontological concepts to give personalized search.

Razmerita et al. [5] presented a generic ontology-based user modelling architecture applied in the context of a Knowledge Management System. Daniela Godoy and Analía Amandi [1] proposed user profiling considering a concept at user interest in varied context and adaptation to changing user interest.

A statistical approach to identify context in a single document is discussed in [11]. Reference [2] proposes a method of keyword suggestion by pre-determined cores which need a lot of pre-processing and core generation from different context documents. [8] and [9] give a different approach by using ontologies for query refinement but they lack in terms of solving ambiguous keywords.

Our proposed system combines a statistical approach along with a method to augment the proposed Context-Ontology to identify the Context of the search. We use the multi-layered Indian-Logic Ontology to represent the ambiguous words in various Contexts. Our approach is a session based approach which treats each and every session of the user as a separate one and it does not store the user preferences permanently. It gives automatic context-oriented query refinement using the first few pages user visits. It visualizes the results based on ontology and gives different Context suggestion for each search, thus it allows the user to completely change the Context of search.

3. Ontology-based query refinement

The overall architecture of 'Context-Aware Automatic Query Refinement Using Indian-Logic based Ontology' [CAQR] is shown in Figure 1. The proposed system automates the process of query refinement without affecting the context of the user-search so that, the user is least disturbed and most satisfied during the search process.

In each user-click a pseudo-feedback is collected automatically from the user clicked pages, based only on user's point-of-view. The pseudo-feedback is the collection of the link address user visits, meta-data and the content of the page along with the original user query. The removal of stop words, stemming are done at the link-analysis level. The backbone of this system is an Indian-Logic Based Context Ontology along with the Indian-Logic Based Domain Ontology. Using the pre-designed Context ontologies the Contextualization process is carried out on the query and the words obtained from the link analysis. This is done by performing a Global and Local analysis to identify the context.

Using the pre-designed Domain ontologies the Conceptualization process is carried out on the query and the context identified from the Contextualization process. This is done by calculating semantic distance [9]. All the concepts semantically related to the context are identified in this process. The optimized query is given to the search engine which retrieves the most relevant pages based on the refined query.

The Re-ranking process is carried out to order the fetched results with respect to the identified domain concepts. As all this processes are iteratively carried so that user gets context-refined pages at each step in every new session and user information is not stored permanently as in the case of personalization.

3.1 Ontology design

The knowledge base of this system is Indian-Logic Based Ontologies classified as Context Ontology and Domain Ontology. Ontology is a formal explicit description of concepts in a domain of discourse. Indian Logic based Ontology is chosen because of its flexibility in structure which helps to depict Concepts in various Contexts. Indian-Logic based Ontology designed using Gautama [6] has the qualities:

- Multi-Layered Concept Representation.
- Each Layer Represents a Context.
- Context is given as an Attribute.
- Attribute is sub-property of Concept.
- Attribute has values.
- Internal relation (concept & attributes).
- Tangential relation (attribute & concepts).
- Grouping relation (attributes & values).
- External relation between any pair of Concepts.

These properties and relations of Indian –Logic Ontology makes it easier in the representation of complex concepts and relations in the multi-layered concept representation proposed for context representation.

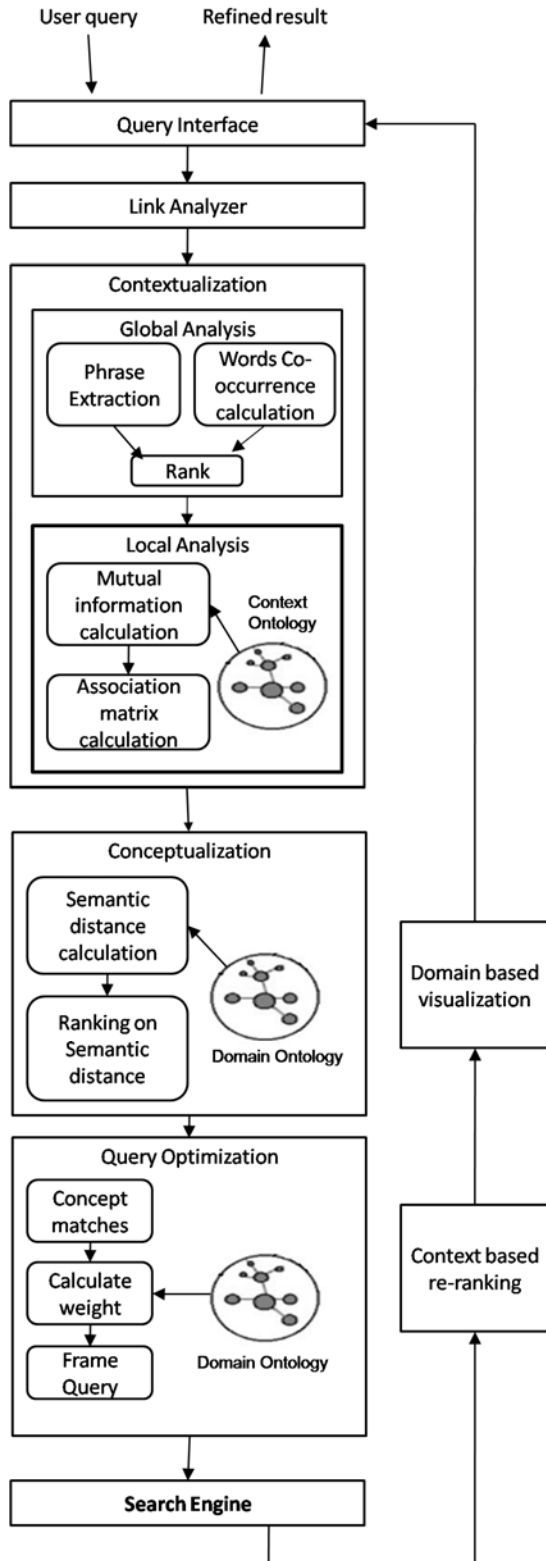


Figure 1. Overall architecture of CAQR system

3.2 Context ontology

Context Ontology gives the structure of ontology proposed to represent concepts in various Contexts. This is achieved using the Multi-layered Concept Representation property of Indian Logic based Ontology. The Context of first layer is considered as the default Context.

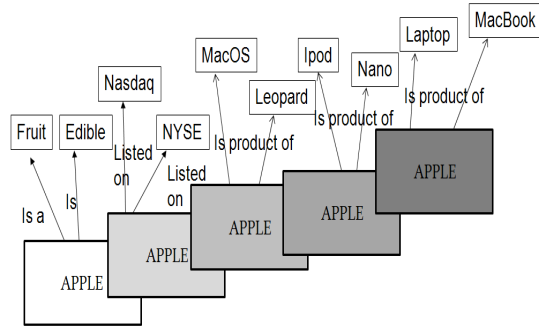


Figure 2. Multi-context representation of concept 'Apple'

The sample concept considered here is 'Apple'. 'Context' is considered as one of the attribute of concept 'Apple'. The attribute 'Context' has distinct values in different layers each representing the concept 'Apple' in different sense as shown in Figure 2. Some relations in this example are given below:

- Grouping relation : Gadgets { IPod, Laptop }
- Tangential relation : Leopard(a) and Animal(C)
- External relation : (MAC OS and PC) Hacking

3.3 Domain ontology

Domain Ontology is used to represent complete data of a domain as a Knowledge base. An example Domain Ontology of "Computer Networks" is shown in Figure 3.

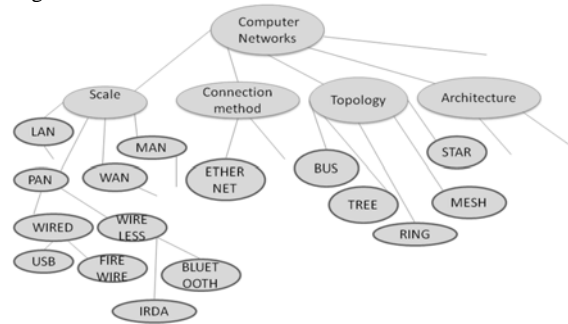


Figure 3. Domain Ontology for "Computer Networks"

3.4 Link analysis

Link Analysis consists of the process of collecting all the data connected with the link that user clicks and the pre-processing of the collected data to remove unwanted data with stopword removal and stemming. The terms related with the user query, terms in headings and terms in metadata are given more weightage.

3.5 Contextualization

Contextualization [Global and Local Analysis] is the core module which helps to avoid the ambiguity in user query by identifying the context of user-search. The context of the user search is identified by analyzing the query against the phrases and co-occurring words extracted from the 'Link-Analysis' word-list using the Indian-Logic based Context Ontology.

3.6 Global Analysis

The basic idea behind this method is from the work proposed in paper [11], which considers the whole textual content, keywords and phrases, extracted using words co-occurrence statistical information to reflect the main theme of the document. Words that collectively forming a phrase give different context to the way they appear alone. So the bi-gram phrases are found by considering all the words and finding the words that precede or follow it in different sentences using the Phrase Extraction algorithm. The original content of the user clicked link page is given as input to the Phrase Extraction algorithm, which identifies and rank phrases on the frequency of occurrence. The higher ranked phrases and co-occurring words reflects the theme of the content by reference [11].

Phrase Extraction Algorithm:

- step1: Let WORD be the first word in total word-list
- step2: Initialize i to zero
- step3: If WORD is a stopword, goto step8
- step4: Set wordpair[i] = WORD
- step5: If i not equal to one, goto step9
- step6: Add wordpair to phrases
- step7: Set wordpair[0] = wordpair[1]
- step8: Set i to -1
- step9: Increment i
- step10: Repeat step3 to step9 for all words in word-list

3.7 Local Analysis

Local Analysis deals with words surrounding the query-word in the sentence in which it occurs. The 'Local Analysis' is by means of calculation of mutual-information and by forming association matrix. This is one of the most used statistical approaches to identify the context of the word based on a single document, proposed in [11] which is enhanced by the addition of a Context -ontology. When a full-fledged Context ontology is available the context suggestions identified using the statistical methods are augmented against the ontology concepts and the correct Context is identified.

3.8 Conceptualization

The basic idea of conceptualization is to provide semantically related suggestion terms to the query which are used to refine the queries. Using the context suggestion from the contextualization module and the domain ontology the domain of the query is identified.

In this module semantic distance is calculated between the queries, contextualization suggested terms

and the domain ontology using the semantic distance measurement proposed in [9]. Based on lowest semantic distance (higher semantic similarity) the semantically relevant concepts are ranked and suggested.

3.9 Query Optimization

The identified Context and Domain Concepts are ranked on the basis of their concept-weight calculated using the number of children and the level of the node in the ontology. The higher ranked term is chosen to expand the query by adding it to the query format for the chosen search-engine to form the refined user query.

The identified Context based keyword suggestions are fed into 'Query Optimizer' which selects the apt keywords for query expansion. The refined query is tailored for the querying format of the selected search engine. The refined-query search results are re-ranked using the identified Context concepts and the domain concepts using the Ontology.

The keyword suggestions other than that selected for query expansion are displayed to the user as recommendations with different possible context combinations, so that the user can change the context of the search or choose keywords from recommendations. The keyword suggestions are grouped into clusters by various relations given in section 3.2 using ontologies.

Keyword Selection Procedure

1. Each node in the ontology is given weightage according to the number of sub classes it has.
2. A class with no subclasses is given weightage as 1.
3. A parent class has twice as much weightage as its subclasses.
4. At each page load the class with the next highest weightage is added to the optimized query.
5. If previously selected class is a parent class then the new class is added to the query.
6. Else the previous class is removed and new class is added.

3.10 Context-based re-ranking

The new search result obtained using refined query is given as input to the Context based re-ranking module. This module extracts the links and their contents in the web-search result page. Feature extraction is done for all the link contents in the result page.

The extracted terms are given to the Context Matcher and the Domain Matcher where all the relevant Concepts in the Context Ontology and the Domain Ontology are identified and based on their weight, the ranker re-ranks the given search result using given algorithm.

Ranking Algorithm:

- step1: Let Cons be list of concepts in identified context
- step2: Let DocFreq be a mapping of documents and frequencies
- step3: Let TermFreq be a mapping of terms and frequencies
- step4: Extract link contents
- step5: Do feature extraction
- step6: if (term not in Concepts) then goto step10

step7: if (term not in TermFreq) then goto step9
 step8: Increment the frequency of the term
 step9: Add (term, 1) to TermFreq
 step10: Repeat step6 to step9 for all terms
 step11: Repeat step3 to step10 for all result links
 step12: Freq = sum of frequencies of all terms in the document
 step13: Add (document.freq) to DocFreq
 step14: Sort DocFreq in descending order of term frequencies
 step15: Display the re-ranked results

3.11 Visualization

The re-ranked result is given to the user as a usual flat list, in addition to it a clustered terms of related context and domain keywords from ontology, linked to the relevant web-link is given to the user.

The Clustering of the relevant keywords in each result link page is done using the ontologies and the visualized result is given as a graph which links the results to relevant keywords, presenting the user with a visual map from which they can select at a glance the sites of most interest to them. It allows the user to easily obtain an overall picture of the types of results returned and also gives indication of relationships and subcategories within results.

4. Implementation and Evaluation

For the ambiguous words in network domain like 'ATM', 'POP', 'MAN' the Google search results are analyzed. The actual and one-time refined queries are tabulated in Table 1. Google result before and after query refinement is compared which shows the improvement by fetching all links related to the intended context and more relevant to the user query. This system improves result by avoiding other unrelated domain pages.

Table 1 Context Refined Queries

Qno	Query	Clicked Link (context)	Refined Query
Q1	man	Network	man + network + ethernet
Q2	man	Human	man + human + male
Q3	atm	Network	atm + network + lan
Q4	atm	Banking	atm + bank + money
Q5	pop	Network	pop + network + email
Q6	pop	Music	pop + music + album
Q7	port	Network	port + network + serial
Q8	port	Harbor	port + harbor + cargo

The CAQR search result compared with actual Google result tabulated in Table.2 shows a significant improvement over the context of the result. The Google actual result for "man" fetched a single result related to network domain citing Wikipedia site. When user clicks the second link depicting "man" in network-domain under CAQR search, and all further searches results are in network-domain and seven out of ten are related with Metropolitan Area Network.

Table 2 Search Result Comparison

Google search result for 'man' vs. 'man+network+ethernet' as on April 2009				
No	Unrefined Results	Context	Refined Results	Context
1.	Man - Wikipedia, the free encyclopedia	Human	Metro Ethernet - Wikipedia, the free encyclopedia	Network
2.	Metropolitan area network - Wikipedia, the free encyclopedia	Network	ethernet and metro ethernet and network Resources TechRepublic	Network
3.	MAN AG: Home	Industry	Ethernet Metropolitan Area Network - What does EMAN stand	Network
4.	Manchester Airport UK (MAN)	Airport	Portland Oregon Ethernet, Internet Access, Metropolitan Area	Network
5.	MAN Nutzfahrzeuge - MAN Truck & Bus UK Ltd	Auto	Ethernet Switched Service - MAN Enterprise Business AT&T	Network
6.	Manchester United Official Web Site	Sports	Ethernet MAN - Telstra Enterprise & Government	Network
7.	BBC SPORT Football Premier League Man Utd dominate PFA	Sports	Metropolitan ethernet network : A move from LAN to MAN	Network
8.	BBC SPORT Football Europe Man City 2-1 Hamburg (3-4)	Sports	Metropolitan Ethernet Network: A Move From LAN to MAN, from ...	Network
9.	MAN Diesel SE - STARTPAGE	Auto	Shedding Light on Optical MAN Services: Ethernet, DWDM and SONET ...	Network
10.	CATHOLIC ENCYCLOPEDIA: Man	Human	Ethernet Network Cards --- Lan Wan Man Broadband & Networking ...	Network

Table 3 Precision Comparison

Query No.	Precision for unrefined results (C2)				Precision for Refined results (C1)			
	Relevant Documents	Total Documents	Precision (%)	Context Match	Relevant Documents	Total Documents	Precision (%)	Context Match
Q1	10	32	31	10	28	32	87	100
Q2	24	32	75	24	28	32	87	100
Q3	8	32	25	8	24	32	75	100
Q4	20	32	62	20	24	32	75	100
Q5	8	32	25	8	24	32	75	100
Q6	15	32	47	15	26	32	81	100
Q7	16	32	50	16	24	32	75	100
Q8	4	32	12	4	18	32	56	100

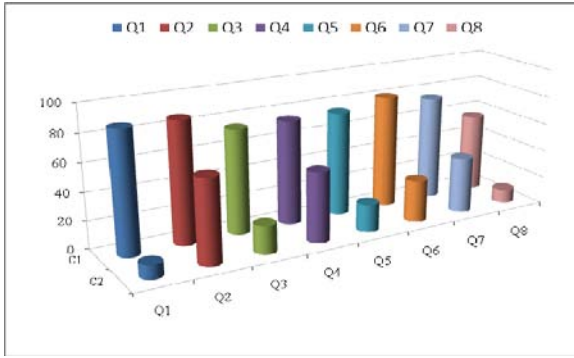


Figure 4. Precision of normal result (C2) vs. CAQR result (C1)

Similar to the first query Q1, all the eight sample queries are analyzed in Table.3 which shows the average Context correctness in Google result is 41% whereas CAQR search has fetched all results in correct context. Figure.4 shows average precision for the eight sample queries for Google result is around 41% considering all the documents it fetched in the correct context are related whereas CAQR search result had a significant improvement in the result by showing 76% precision in the first iteration of refinement. Since this is an iterative refinement the precision will be increasing gradually if the user clicks on the relevant results in each of the search.

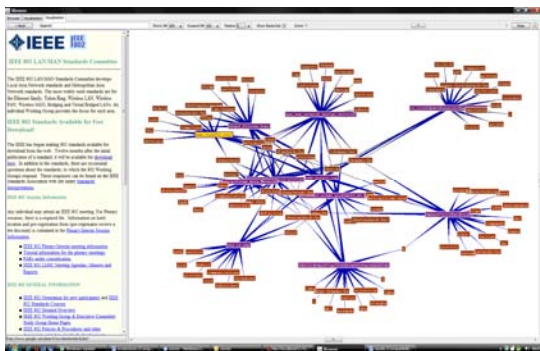


Figure 5. Visualization of the refined search result

The visualization of the refined result for keyword 'MAN' is shown in Figure.5 which visualizes the keyword as the centre point and relates the first eight results as links to the keyword showing each links with a cluster of keywords matching the domain concepts in ontology. Clicking on any of the keywords will open the result in the side panel highlighting the clicked keyword. This visualization gives overall picture of the types of results obtained and also gives indication of relationships and subcategories within the refined results.

5. Conclusion and Future-enhancements

This System performs a contextualization and conceptualization on the user clicked web-pages to extract information to disambiguate the query term and

identify the context of the query terms using the proposed Indian-Logic based Context-Ontology. Hence this system provides a Context aware query refinement to fetch web-pages more relevant to the user query.

This system in the tests provided a convincing improvement over the actual Google result by fetching pages all related to the context intended by the user thus improves precision substantially. For the actual query the default Google result page contained only a single result in Network Context whereas the refined result contained all top-ten results related to the network Domain. Since this system depends on Ontology, the automatic generation of Ontology and prediction of user intended change in context will be considered as future enhancements.

6. References

- [1] Daniela Godoy and Analía Amandi, "User Profiling for Web Page Filtering", *Internet Computing*, IEEE, vol.9, no.4, July 2005, pp. 56-64.
- [2] Eduardo H.Ramirez, Ramon F.Brena, "Semantic Contexts in the Internet", *Fourth Latin American Web Congress (LA-WEB-)*, IEEE, October 2006, pp. 74-81.
- [3] Jens Graupmann, Jun Cai and R. Schenkel, "Automatic Query Refinement Using Mined Semantic Relations", *International Workshop on Challenges in Web Information Retrieval and Integration*, April 2005, pp. 205-213.
- [4] Joana Trajkova, and Susan Gauch, "Improving Ontology-Based User Profiles", *RIAO Conference*, April 2004, pp. 380-389.
- [5] Liana Razmerita, Albert Angehrn, and Alexander Maedche, "Ontology-based User Modeling for Knowledge Management Systems", *International Conference on User Modeling*, June 2003.
- [6] G. S. Mahalakshmi, T. V. Geetha, Arun Kumar, Dinesh Kumar, S. Manikandan, "Gautama --- Ontology Editor Based on Nyaya Logic", *Third Indian Conference on Logic and Its Applications (ICLA) LNAI 5378*, Springer-Verlag, 2009, pp. 232-242.
- [7] Maria Golemati, Akrivi Katifori, Costas Vassilakis, George Lepouras, and Constantin Halatsis, "Creating an Ontology for the User Profile: Method and Applications", *International Conference on Research Challenges in Information Science*, IEEE, 2007.
- [8] Nenad Stojanovic, "An approach for ontology-enhanced query refinement in information portals", *International Conference on Tools with Artificial Intelligence (ICTAI 2004)*, IEEE, November 2004, pp.346-357.
- [9] Rajasekar S.N, Mahalakshmi G.S, and Sendhilkumar .S "Ontology based Automatic Query Refinement", *International Journal of Artificial Intelligence and Soft Computing*, Inderscience Publishers, vol 1, no 2-3, 2008.
- [10] Susan Gauch, Jason Chaffee, and Alexander Pretschner, "Ontology-Based User Profiles for Search and Browsing", *User Modelling and User Adapted Interaction (UMUAI)*, The Journal of Personalization Research, Special Issue on User Modelling for Web and Hypermedia Information Retrieval, 2003.
- [11] Yueqin Hang, Jie Shen, Yin Lin, and Zhao Minya, "Context Information Extraction of the Query Based on Single Document", *International Conference on Services Computing (SCC'04)*, IEEE, 2004.

This article was featured in

computing **now**

ACCESS | DISCOVER | ENGAGE

For access to more content from the IEEE Computer Society,
see computingnow.computer.org.



IEEE  computer society

Top articles, podcasts, and more.



computingnow.computer.org