Providing Support Features for Novice Researchers using Semantic Approach
Maizatul Akmar Ismail, Mashkuri Yaacob, Sameem Abdul Kareem & Fariza Hanum Nasaruddin
Faculty of Computer Science and Information Technology
University of Malaya
50603 Kuala Lumpur,
MALAYSIA

Abstract:
The access to internet has tremendously changed the way of information dissemination. The emergence of digital libraries and institutional repositories provides endless supplies of knowledge. Scholars in particular, make use of research output in the form of conference proceedings, journal and theses as references as guideline in generating new knowledge for the use of future generations. On the other hand, novice researchers are the form of scholars which always drown in the ocean of information. Support in the early stage of study is crucial for novice researchers as it will give them some insights of where to seek for extra information based on institutions, people and research trend without having to go through tedious process of identifying this information all by themselves. Previous studies have identified support features that are useful for novice researchers which among others are Relevant Literatures and Expert Detection. Thus, the purpose of this paper is to explore each of these support features and suggest the state-of-the-art development approach by utilizing Semantic Web technologies. The algorithms involved with each of the support features will be discussed. The result of the implementation shows significant information that can be utilized by novice researchers in accelerating research process. Future work in enhancing the proposed prototype will be discussed as concluding remarks.

Key-Words: Novice Researchers, Support Features, Semantic Approach

1 Introduction
The web provides huge amount of information which offers unlimited resources for research work. On the other hand, it provide extreme amount of search results that may cause the scholars to be lost in cyber ocean. Support in the early stage of study is crucial for novice researchers as it will give them some insights of where to seek for extra information based on institutions, people and research trend without having to go through tedious process of identifying this information all by themselves.

The production of academics activities from Faculty of Computer Science and Information Technology, University of Malaya (FCSIT, UM) are mostly reflected in the Malaysian Journal of Computer Science (MJCS), Malaysian Journal of Library and Information Science, locally developed Conference database and collection of theses stored in FCSIT DSpace Institutional Repository. These resources are heterogeneous in nature, and need to be integrated. Semantic approach has been chosen as an integration method which have been described in (Ismail et al., 2008a, Ismail et al., 2008b). Support needed for research work have been discussed widely in (Kampa, 2002), (Ismail et al., 2008c) and (Ismail et al., 2007). This paper is going to discuss the methodology taken in realizing two of the support features identified in (Ismail et al., 2008c) by utilizing semantic approach.

2 Methodology
2.1 Scholarly Documents Preparation
We utilized Ontogen, a semi-automatic ontology editor in assigning relevant topics to the scholarly documents such as thesis, conferences and journal articles (Fortuna et al., 2006). Experts in four areas of Computer Science were asked to relate the documents with its suitable topics in the ontology. The topics ontology is constructed based on the research category in ACM and the categories that are suggested by the authors of the papers. For instance, the concepts and terms in the following table are extracted from the Malaysian Journal of Computer Science and Malaysian Journal of Library and Information Science (see table 1.0).
These journals are chosen because of the high recognition received from international bodies for scientific community i.e. SCOPUS and International Science Index (ISI). Furthermore, these journals reflect the publication of high quality papers by majority of local researchers. The table shows the answer given by one of the expert based on the relatedness of topics in Information System (row) and Information Science (column). Relevant terms for each concept are supplied based on the keywords assign by the author of the documents.

Table 1.0 : The cross relation of concepts from Information Science and Computer Science

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Terms</th>
<th>Synonyms (Equivalent Words)</th>
<th>Information System</th>
<th>Information Storage and Retrieval</th>
<th>Database Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Applications in Information Science</td>
<td>Online Catalogs</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Portal Technologies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Common User Interface</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-commerce</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet Technologies</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Digital Libraries</td>
<td>Electronic Journals</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Electronic Resources</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Resource Assessment</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>E-Publishing</td>
<td>Online Information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-Book</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Information Management</td>
<td>Business Process Reengineering</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Information Services</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>E-Faculty</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electronic Records</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-Governance</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.0 : Ontogen’s environment in constructing domain ontology for Information System and Information Science.

2.2 Relevant Literature Module

Identification of relevant literature is identified by research expert to be the most important support that would assist novice researchers in the early days of their studies. The Related Literature feature will retrieve information through a matching module which supported by a domain ontology such as ACM Computing Classification System (1998) and individually developed ontology based on research category suggested by the author. An ontology based semantic layer would enable the relationships between the various metadata schemes to be properly represented within the chosen ontology. The ability of using relationship information which is stored in ontologies enables semantic search engines to overtake the problems associated with existing search methods. The result of search will be the literatures based on the most related one to the ACM category and Ontogen’s constructed ontology with the entered keywords. This is beneficial because accurate categorization will provide quick content reference for the researcher and accelerate the progress of the search for related literature, as well as searches for the research on other online resources.

The algorithm that we applied for relevant literature module is as follows:
1. User Input Query
2. Extract Abstract, Topic and Keywords (User assigned keywords and Ontogen’s suggested keywords) from triple store database.
3. Create an empty Array
4. Tokenized Abstract, Topic and Keyword (User assigned keywords and Ontogen’s suggested keywords)
5. Add each Token to Array
6. Find number of user query in Array of Tokens
7. Rank and List Articles and Category based on frequency
8. Bring all other articles that are attached to sub-category and super-category of ontology
9. Bring all other articles that are attached to sub-category and super-category of ontology

2.3 Expert Detection Module
In detecting expert(s) in the field of CS and IT, the text corpus will be analyzed by using statistical technique to find specific experts in specific research field. Typically, these are some questions that novice researcher has in mind in order to justify other researchers’ level of expertise:
- This author has any other papers?
- Who are the experts in digital libraries?

Resolving those queries manually would require further investigation and research which would requires more time. In determining the expert in our work, we used three heuristics as below (Table 2.0):

<table>
<thead>
<tr>
<th>No.</th>
<th>Parameters</th>
<th>Weightage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of Publications</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>Citation Analysis (Published papers in (1) that have been cited by others)</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>Sequence of Authors</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Total ((1) + (2) + (3))</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2.0 : Heuristics Used in Determining Experts

Bibliographic measures are widely known in measuring the performance of the members of scientific community (Katz and Hicks, 1997). Numbers of publications in specific research area have long been recognized as indicators of expertise (Kampa, 2002, Kampa and Carr, 2000, Katz, 2000). Thus, in determining the expertise, size of published papers is given more weight compared to other indicators that are described by Katz (2000) i.e. citation of papers and citations per paper. Impact factors for published papers is measured by the citations that are received by particular papers (Yinian and Zainab, 2001). There are few techniques in determining author impact factors. First, the h-indexed formulae is applied to measure the impact factors (Hirsch, 2005) : "A scientist has index h if h of his/her Np papers have at least h citations each, and the other (Np – h) papers have no more than h citations each.". For instance, h-index of 10 means that a researcher has published 10 papers that each had at least 10 citations. Second, cited references or citation analysis involves “counting how many times a paper or researcher is cited” in which concluded that “influential scientists and important works are cited more often than others”(HealthLinks, 2008). Google Scholar and Thomson ISI Web of Science are often used for citation analysis (Harzing, 2008). Consequently, we assigned citation analysis weight lower than the number of published papers i.e. 40%. Other indicator of expertise that is known by the academic community but seldom be given any attention is the order of authors’ appearance in academic papers. (Subramaniam, 1983) stated that “The question of ordering the names of coauthors is highly complex and elusive, while it is generally true that the name of the principal investigator is almost always mentioned first... “(p.36). Even though the name of co-authors are sometimes arranged according to alphabetical order, the principals investigators name whom contributed the most for the research publication is almost placed at the beginning (Zuckerman, 1968). Thus, we take this heuristic into consideration and give the weightage of 10% in determining expertise in particular research area. As (Subramaniam, 1983) observed that it is not uncommon to find ten or more names for research publications, random analysis on ACM and IEEE publication from the year 2000 onwards shows that the most authors in a single publication are nine. Therefore, we put the maximum weight for the first author weight and the least weight for the last author. Researchers often involved in various research
works. For this, we also suggest the super-category based on ACM’s and Ontogen’s classification with the assumption that the super-category of research topics will give the user some idea on the people that are working in the similar area of research. We take into account the nearest subcategory and supercategory of research area in ranking other experts that work in similar area of research. Figure 2.0 below shows the ontology Information System’s categories.

For instance, if the user entered Information Retrieval as query and want to see the experts working in this area, we applied the three heuristics that we have mentioned above (see table 2.0). We also suggested other experts that work in similar area based on the supercategory of Information retrieval and rank these experts according to the distance of nodes. The depth of nodes in our data is four, thus we assign most weight to the nearest nodes and least weight to the nodes that are far from the user query. As the research topics are suggested by the research experts in the category, this technique is considered reliable because novice researchers can have more options in selecting other expert researchers in guiding them in the research work in case that the chosen topics do not have any experts in it. The algorithm that we applied for expert detection module is as follows:

1. User Input.
2. Search inside ontology for the articles in the query category.
3. Extract the articles from triple store database.
4. Keep result in the array
5. Find the ID of the articles to search its title.
6. If Array != NULL
   a. Go to Step 7
   b. Else go to Step
7. Extract the titles.
8. Search for that title in the Reference database to see who cited it & if it is cited more than two times – assign 0.4 weights (1).
9. Check Sequence of Authors (2) - assign weight (0.09-0.1).
10. Extract the authors of those Articles.
11. Keep Authors in an array.
12. If Array != NULL
    a. Go to Step 13.
    b. Else go to Step
13. Compare the authors to find which one repeated more.
14. Store the authors name and articles ID and (3) number of repeat the array – assign 0.5.
15. Sort using (1) + (2) + (3).
16. Show List of articles.
17. Find tree of super category.
18. Is it the nearest category?

3 Conclusion
Most of the important steps in identifying the right information at the right place could be done faster if research students have been told to do the “right” things first. Certainly, it does depend on the students’ effort to ascertain the success of their research. Thus, extra support on the earlier stage of the study would definitely be helpful. This paper highlight two types of support features needed to assist novice researcher and the steps taken in realizing it. Future work involves the development of other features identified in (Ismail et al., 2008c) i.e. Trend Detection module, Relevant Online Database module and Relevant Research Institution module.

References:
Sciences Libraries.


