DBLP-SSE: A DBLP Search Support Engine

Yi Zeng
International WIC Institute
Beijing University of Technology
Beijing, P.R. China 100124
yzeng@emails.bjut.edu.cn

Yiyu Yao
Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
Email: yyao@cs.uregina.ca

Ning Zhong
Department of Life Science and Informatics
Maebashi Institute of Technology
Maebashi-City, 371-0816, Japan
zhong@maebashi-it.ac.jp

***Abstract*—A Search Support Engine (SSE) is implemented based on the basic principles of Information Retrieval Support Systems (IRSS) and Information Seeking Support Systems (ISSS). An SSE aims at meeting the diversity needs from different users, providing various supporting functionalities, tools, etc. for users to perform various tasks beyond the traditional search and browsing provided by current search engines. As an illustrative example, we developed a DBLP search support engine (DBLP-SSE), and we discuss some concrete supporting functionalities, namely, search refinement support, domain analysis support, etc. Each of the functionality focus on a unique perspective supporting users finding useful information and knowledge from the DBLP dataset. The search support engine can be considered as a step towards Knowledge Retrieval (KR) and Web Intelligence (WI).**

***Keywords*-search support engine; information retrieval support system; information seeking support system; user interest modeling; knowledge retrieval**

I. Introduction

With the advances of search engines and public data on the Web, everyone has to face a great challenge of finding useful information and knowledge. One is no longer satisfied with traditional browsing and search. There is a great demand on search engines with much more intelligence and new functionalities to support search [11]. Information Retrieval Support Systems (IRSS) was proposed to provide more functionalities to support users get what they need [13]. As one type of concrete systems that implements the basic principles of Information Retrieval Support Systems (IRSS) [13] and more recent developed Information Seeking Support Systems (ISSS) [9], we suggest a Search Support Engine (SSE). In order to satisfy various needs from different users, an SSE provides various supporting functionalities, tools, and allows them to perform various tasks beyond the traditional search and browsing functionalities which are provided by the current search engines.

As a concrete search support engine on the Web, we developed a DBLP search support engine (DBLP-SSE), which focuses on providing more support functionalities for users to use the DBLP [7]. In this paper, we will introduce some concrete supporting functionalities provided by the system. Namely, search refinement support, domain analysis support, etc. Each of the functionality focus on a unique perspective supporting users finding useful information and knowledge from the DBLP system. The search support engine is aimed at a step towards Knowledge Retrieval (KR) [14] and Web Intelligence (WI) [15].

The paper is structured as follows. Section II summarizes the related principles and components that support the development of search support engines. Section III introduces the DBLP-SSE system, its main functionalities, theoretical basis, and some concrete implementation examples. Section IV concludes the paper by highlighting the major contributions.

II. Basic Principles and Architecture of Search Support Engines

To meet the Web search needs, Search Engines (SE) were proposed and developed based on the principles of Information Retrieval (IR) [3], the development of Search Support Engine (SSE) is based on the principles of Information Retrieval Support Systems (IRSS) [13]. By moving beyond browsing, navigating and retrieval, IRSS focus on a wide range of supporting functionalities, including summarization, exploration, analysis, knowledge discovery, results organization, and so on [13], [6]. Recently, information seeking support system (ISSS) emerged with similar aim [9]. Taking the philosophy of IRSS, SSE not only provide component for traditional navigation, search and browsing, it also provide many other supporting components, For example, knowledge organization, knowledge discovery, knowledge visualization components, and many more. Instead of emphasizing the search functionality, SSE emphasizes providing users with different supporting functionalities.

Many existing studies begin to add supporting functionalities to search engines, but they are typically implemented by simply adding components or plugins on the traditional architecture of search engine. There is still a lack of basic architecture for SSE. Adding supporting functionalities on current architecture of search engine does not create a promising SSE. For this reason, we developed a layered architecture of SSE, as shown in Figure 1. The main functionalities for the major subsystems are listed as follows:

- User Interface Subsystem : provides a friendly interface for user system interaction (connecting users, the

User Interface Subsystem			
Search Subsystem		Search Support Subsystem	
Data Management Subsystem	Knowledge Base Subsystem	User Profile Management Subsystem	...

Figure 1. A layered architecture of search support engine.

search subsystem and the search support subsystem).

- Search Subsystem : performs the search and inference tasks, and analyzes search results and provides the results to users through user interface.
- Search Support Subsystem : contains different functionalities, and tools for search support and post processing.
- Data Management Subsystem : collects and manages data, information, and knowledge. It can be a database system, file system, etc.
- Knowledge Base Subsystem : manages various domain knowledge for search support.
- User Profiles Management Subsystem : manages user profiles to support the diversity of search support needs.

III. DBLP-SSE

In order to support various search functionalities on the DBLP Computer Science Bibliography [7], many systems have been developed. Such as FacetedDBLP, CompleteSearch, DBLPVis, etc. FacetedDBLP divides search results according to topic facets (such as by conference, year, etc.) [5]. CompleteSearch provides instant response after each keystroke, prefix search, categorized results, etc [4]. DBLPVis provides visualizations of relations between authors, publication sources and terms [10]. All of the mentioned systems can be considered as search support engines based on DBLP that extend some of the search functionalities. In this section, we provide a system that is different from other systems and focuses on some extra functionalities.

A. Main functionalities of DBLP-SSE

In DBLP-SSE[1], we mainly provide two types of supporting functionalities, namely, search refinement support and domain analysis support. For the search refinement support, DBLP-SSE first track the change of each authors research interests and make a prediction of his/her current research interests based on some interest retention models, then it use acquired user interests as implicit query constraints to refine incomplete or vague queries from users. Through this supporting functionality, search results which are consistent with predicted user interests are ranked into the top ones, and users can easily find the results which may be most relevant to their needs although they may not explicitly put enough necessary constraints to the input query. For the domain analysis support, DBLP-SSE provides support on building domain structures, tracking domain trend, finding author distributions, etc. These functionalities are also user centered, if the users log on the system, based on predicted user interests, the system can automatically generating relevant domain analysis results to users so that they can be aware of the change in this domain.

[1]More supporting materials can be found through the DBLP-SSE website http://www.iwici.org/dblp-sse

B. Theoretic basis

In order to provide mentioned functionalities, many theoretical basis are needed. In this paper, we focus on introducing the user interest modeling and prediction methods. For the purpose of providing better knowledge query results, we extract all the authors' publication lists and analyze the publication history, then we predict current research interests based on the models proposed in this section and use acquired interests as implicit query constraints for the original query.

For simplicity, we provide a model of measuring research interests by counting keywords or terms appearing in all publications based on the following equation:

$$TRI(i) = \sum_{j=1}^{n} m(i,j), \tag{1}$$

where $j \in [1,n]$ is the number of years involved, $m(i,j)$ is the number of appearances of term i in the year j, $TRI(i)$ reflects the value of total research interest on topic i, namely, how many times has a research interest appeared in the considered years. The above computation may not correctly reflect a researcher's current research interests. For example, he/she has shifted the research interest, but the accumulated publications in an old research field may still be higher than that in a new field.

Research interests are to some extent similar to memory in cognition. The loss of interest in an area can be regarded as forgetting of a previously remembered knowledge. We propose a way to model the retention of research interests ($RRI(i)$) based on memory retention. Memory retention models can be categorized into two types, that based on exponential law [8] or that based on power law [12].

The exponential model suggests that the forgetting function satisfies an exponential formula $P = Ae^{-bT}$, where P represented the performance measure of memory retention, A and b were two parameters for the model, and T was the period of time remembered (delay time) [8], [2]. Based on the exponential formula, the retention of a research interest can be denoted as:

$$E\_RRI(i) = \sum_{j=1}^{n} m(i,j) \times Ae^{-bT_i}, \tag{2}$$

where we consider the time remembered T_i for a topic i in a yearly manner. For each year j, there might be

$m(i,j)$ publications on a specific topic i, and each of them will contribute a value Ae^{-bT_i} to the total retention of a research interest contributed by that year, where A and b are constants. $E\_RRI(i)$ is the retention of a research interest i through all the years based on the exponential model, and is very related to the current interests for a researcher.

Another model for retention of research interests can be represented based on a power function for memory retention [12], [2]:

$$P\_RRI(i) = \sum_{j=1}^{n} m(i,j) \times AT_i^{-b}, \tag{3}$$

where T_i is the number of years interested in topic i until a specified year, $m(i,j)$ is the number of publications on a specified topic i in the year j, and $m(i,j) \times AT_i^{-b}$ is the total retention of a research interest contributed by that year.

To sum up, $TRI(i)$ reflects an author's interest on topic i through all the counted years, which reflects the total interest value through an author's academic life. $E\_RRI(i)$ and $P\_RRI(i)$ focus on the interest retention on the topic i in more recent years, hence they can be used for prediction of current interests.

C. Experiments

1) User Interests Support: As an illustrative example, we consider a scenario of tracking the authors' research interests, which are implicitly embedded in their own publication lists. There are many literature search systems which provide author publication lists (e.g. Web of Science, CiteSeerX, PubMed, Google Scholar, etc.). In our study, we use the SwetoDBLP dataset [1].

In our study, in order to minimize the value of ρ in t-test, the parameters in the power law model are chosen as $A = 0.855$ and $b = 1.295$. The value for Spearman's rank order correlation coefficient between the prediction and the real data is $\gamma \approx 0.107$, and for 1-tail t-test, $\rho = 0.237$. For the exponential law model, the parameters are chosen as $A = 0.535$ and $b = 0.382$. The rank correlation coefficient is $\gamma \approx 0.168$, and for 1-tail t-testing, $\rho = 0.129$. The results are, to some extent, close to statistical significance. In order to test the parameters in larger range, in our initial work, we choose all the authors from the SwetoDBLP dataset whose number of publications are above 100 (1226 authors in all). Using power functions and relevant parameters introduced above, we extract top 9 interests from their interest lists from the year 2000 to 2007 (hence, 1226×8 sets of data are involved). A comparative study on the actual interests and predicted interests has been done. According to the experiment results, 0.98% of the prediction can match at least 7 interests, 3.22% can match 6 interests, 8.35% can match 5 interests, 15.66% can match 4 interests, and 21.33% can match 3 interests. Hence, in all, 49.54% of the predictions can match at least 3 interests in the top 9 interests in our experiment. From this experiment result, we can conclude that, to some extent, the prediction based on interests retention has some degree of relation with the actual publication numbers, even though there is some gap from statistically significant for the test of Spearman's rank order correlation coefficient between the model prediction results and the real data.

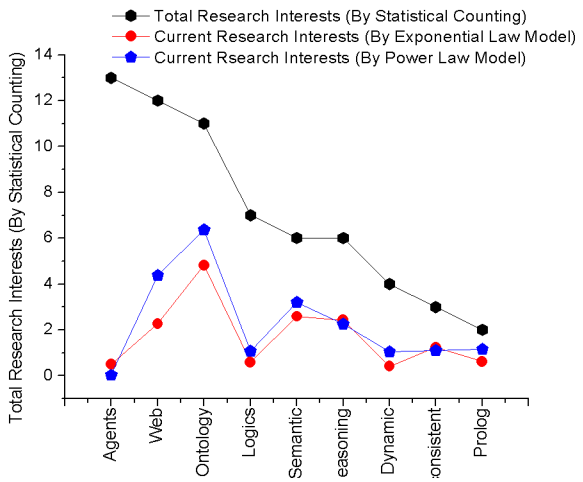

Figure 2. A comparative study of total research interests through the years 1990-2008 and current research interests in 2009 (based on both the power law and exponential law models)

Table I
A COMPARATIVE STUDY OF TOTAL RESEARCH INTERESTS AND CURRENT RESEARCH INTERESTS (2009) OF ZHISHENG HUANG BASED ON THE DBLP PUBLICATION LIST

TRI	P_RRI	E_RRI
Agent(13)	Ontology(5.9041)	Ontology(4.8218)
Web(12)	Web(4.5450)	Semantics(2.5867)
Ontology(11)	Semantics(3.0551)	Reasoning(2.4257)
Logic(7)	Reasoning(2.4845)	Web(2.2742)
Semantic(6)	Prolog(1.2034)	Inconsistent(1.2383)
Reasoning(6)	Inconsistent(1.2672)	Prolog(0.6143)
Dynamic(4)	Logic(1.2567)	Logic(0.5847)
Inconsistent(3)	Dynamic(0.9889)	Agent(0.4921)
Prolog(2)	Agent(0.8741)	Dynamic(0.4112)

A comparative study of Zhisheng Huang's total research interests and current research interests (through the values of interest retention by an exponential law ($A = 0.535$ and $b = 0.382$) model and a power law ($A = 0.855$ and $b = 1.295$) model) are shown in Tables I, and Figures 2, respectively. By using these two models, users' background information can be obtained for further use in the search process.

From Table I, we can observe that for some research interests, even if they have a big value on the total research interest through Zhisheng Huang's research life (up to now), they may not be his current major research interests. Taking "Agents" as an example, it has the highest value of total research interests, but has very low current research interest based on the computation of his research interests retention. Although it is in the third place of the total research interests, "Ontology" is the number 1 current research interest based

on both the power law model and the exponential law model.

We developed a user interests support component, which extracts author names, publication lists and corresponding years from the SwetoDBLP dataset [1], version April 2008, and reports their current interests based on the introduced power law model and the exponential law model. The important issue in here is that how to utilize the current user interests prediction to support various functionalities provided by the DBLP search support engine.

2) Search Refinement Support: For incomplete or vague query on the Web, the number of retrieved results for user inspection can be very huge. Furthermore, if the query cannot describe user needs as precisely as possible, it might be very hard to obtain a good set of search results. It is not surprising to predict that in most cases, user query are very relevant to his interests. Hence in DBLP-SSE, we provide a search refinement support component based on query extension using user interests acquired from the user interest support component.

The search refinement support component allows an author to log in using his/her own name which is consistent in the DBLP publication list, then the system will generate a series of current interest keywords from his/her own publication list. When the user inputs a query in the search box, the system will automatically add constraints from the user interest list, and two lists of search results are provided. The first list is produced using the original query, and the second is produced using the refined query which includes the top 9 interest keywords from the user interest list. A screen shot of the system and a case study are provided in Figure 3 and Table II. After logging in using the name Dieter Fensel, the user provides a query input, and different results are presented according to the original query and the refined query. As we can see, list 2, is much closer to the user's interests. In this way, most relevant search results to the specified user are ranked to the top ones.

3) Domain Analysis Support: In domain analysis support, here we introduce two types of supporting functionalities, namely, building domain structures from publication lists and finding author distributions.

Building Domain Structures

Figure 4 shows a domain structure of the field "Artificial Intelligence", which is based on an analysis of proceedings indexes of the 1969-2007 International Joint Conferences on Artificial Intelligence (IJCAI). We can infer that if one needs very general information with respect to the field "Artificial Intelligence," he/she may just want the knowledge in the second level (the first level just has one node "Artificial Intelligence"), which includes around 100 branches of AI (in fact, if we do not organize the index of these proceedings into a hierarchical knowledge structure, one can get around 400 branches which are very confusing in one level). Furthermore, if he/she needs more detailed knowledge concerning one branch of AI, say "Robotics", he/she can choose "Robotics" and get a finer grained structure. In this way, we can produce a scalable knowledge structure which provides the knowledge source in different levels of details with an interactive manner concerning different user needs.

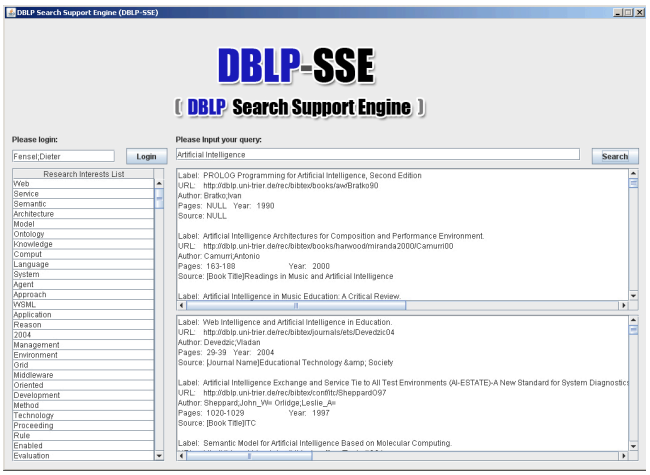

Figure 3. Search refinement support using DBLP-SSE.

Table II
A COMPARATIVE STUDY OF SEARCH RESULTS FROM THE ORIGINAL QUERY AND THE EXTENDED QUERY WITH USERS' CURRENT INTERESTS

Name	Dieter Fensel
Top 9 interests	Web, Service, Semantic, Architecture, Model, Ontology, Knowledge, Computing, Language
Query	Artificial Intelligence
List 1	without current interests constraints
	* PROLOG Programming for Artificial Intelligence. * Artificial Intelligence Architectures for Composition and Performance Environment. * Artificial Intelligence in Music Education: A Critical Review. * Music, Intelligence and Artificiality. Artificial Intelligence and Music Education. * Musical Knowledge: What can Artificial Intelligence Bring to the Musician? * .....
List 2	with current interests constraints
	* Web Intelligence and Artificial Intelligence in Education. * Artificial Intelligence Exchange and Service Tie to All Test Environments (AI-ESTATE)-A New Standard for System Diagnostics. * Semantic Model for Artificial Intelligence Based on Molecular Computing. * Open Information Systems Semantics for Distributed Artificial Intelligence. * Artificial Intelligence and Financial Services. .....

Finding Author Distribution

The author distribution in each branch of a field may be useful to some authors, based on which the authors could find potential coauthors who are with similar interests. Figure 5 provides an example of author distribution in some fields of Artificial Intelligence. The number of authors in each field is based on term search on the publication list.

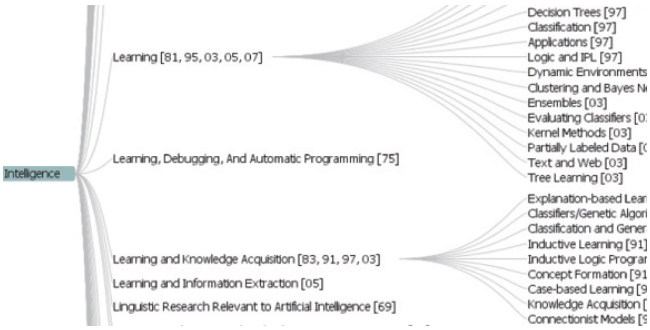

Figure 4. A partial multi-level knowledge structure of Artificial Intelligence according to analysis on proceedings indexes of IJCAI 1969-2007.

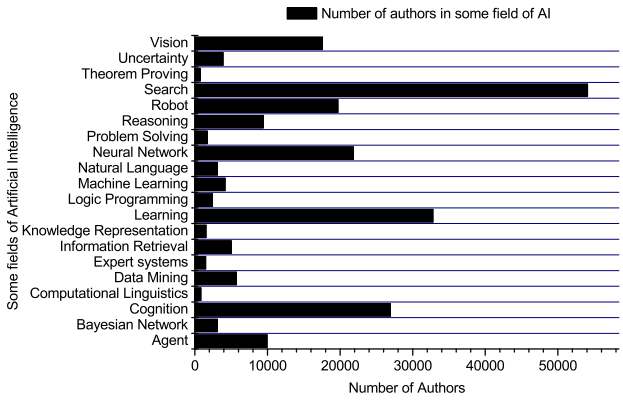

Figure 5. Author distribution in some fields of Artificial Intelligence

IV. Conclusion

Search support engine is introduced to meet the diversity needs from various users based on Web scale data through providing more supporting functionalities that go beyond traditional search and browsing. The theoretic basis introduced for user interest modeling can be applied to other areas, such as user interests retention for general Web search. More supporting functionalities will be added to future versions of the DBLP-SSE. Experimental results have shown some potential impact that search support engine may be considered as a step towards Web Intelligence [15].

Acknowledgment

This work is supported by the European Union 7th framework project Large-Scale Integrating Project, Large Knowledge Collider (LarKC). Special Thanks to Yang Gao and Yan Wang who has participated in system development, and to Zhisheng Huang, Lael Schooler and Jose Quesada for their detailed comments on some previous work. The authors would like to thank anonymous reviewers on their constructive comments.

References

[1] Aleman-Meza, B. Hakimpour, F., Arpinar, I., and Sheth, A. Swetodblp ontology of computer science publications. *Journal of Web Semantics 5*, 3 (2007), 151–155.

[2] Anderson, J., and Schooler, L. Reflections of the environment in memory. *Psychological Science 2*, 6 (1991), 396–408.

[3] Baeza-Yates, R., and Ribeiro-Neto, B. *Modern Information Retrieval*, 1 ed. Addison Wesley, New York, 1999.

[4] Bast, H., Mortensen, C., and Weber, I. Output-sensitive autocompletion search. *Information Retrieval 11*, 4 (2008), 269–286.

[5] Diederich, J., Balke, W.-T., and Thaden, U. Demonstrating the semantic growbag: Automatically creating topic facets for faceteddblp. In *Proceedings of the ACM IEEE Joint Conference on Digital Libraries* (2007).

[6] Hoeber, O. Web information retrieval support systems: The future of web search. In *Proceedings of the 2008 International Workshop on Web Information Retrieval Support Systems* (2008), vol. 3, pp. 29–32.

[7] Ley, M. The dblp computer science bibliography: Evolution, research issues, perspectives. In *Proceedings of the 9th International Symposium of String Processing and Information Retrieval* (2002), pp. 1–10.

[8] Loftus, G. Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory and Cognition 11* (1985), 397–406.

[9] Marchionini, G., and White, R. Special issue on information-seeking support systems. *IEEE Computer 42*, 3 (2009), 30–66.

[10] Pohl, M., Reitz, F., and Birke, P. As time goes by: integrated visualization and analysis of dynamic networks. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (2008), pp. 372–375.

[11] Solso, R., MacLin, O., and MacLin, M. *Cognitive Psychology*. Allyn & Bacon, Boston, Massachusetts, 2007.

[12] Wickelgren, W. *Handbook of learning and cognitive processes*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1976, ch. Memory storage dynamics, pp. 321–361.

[13] Yao, Y. Information retrieval support systems. In *Proceedings of the 2002 IEEE International Conference on Fuzzy Systems* (2002), pp. 773–778.

[14] Yao, Y., Zeng, Y., Zhong, N., and Huang, X. Knowledge retrieval (KR). In *Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence* (2007), pp. 729–735.

[15] Zhong, N., Liu, J., and Yao, Y., Eds. *Web Intelligence*, 1 ed. Springer, 2003.