Social Relation Based Scalable Semantic Search Refinement

Yi Zeng[1], Xu Ren[1], Yulin Qin[1,2], Ning Zhong[1,3], Zhisheng Huang[4], Yan Wang[1]

[1] International WIC Institute, Beijing University of Technology
Beijing, 100124, P.R. China
yzeng@emails.bjut.edu.cn

[2] Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213, U.S.A
yq01@andrew.cmu.edu

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

[4] Department of Artificial Intelligence, Vrije University Amsterdam
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
huang@cs.vu.nl

Abstract. One of the major problems for semantic search at Web scale is that the search results on the semantic data might be huge and the users have to browse to find the most relevant ones. Plus, due to the reason for the context, user requirement may diverse even the input query may be the same. In this paper, we try to achieve scalability in semantic search through social relation diversity of different users. Namely, we utilize one of the major context for users, social relations, to help refining the semantic search process. Social network based interest retention model is developed on top of user name based social relations, and is designed to be used in more wider range of Web scale semantic search tasks. The experiments are based on the SwetoDBLP dataset, and we can conclude that proposed method is potentially effective to help users find most relevant search results in a scalable environment.

Keywords. social relation, interest retention, social network based interest retention model, semantic search, search refinement.

1 Introduction

Formulating a good query for search is an everlasting topic in the fields of information retrieval and semantic search, especially when the data comes to Web scale. The hard part is that users some times cannot provide enough constraints for a query since many of the users are not experienced enough. User background is a source that can be used to find user interests and the acquired interests can be added as constraints to the original vague query to refine the query process and help users get most relevant results.

In our setting for this study, we define a user interest as concepts that the user is interested in or at least familiar with. In addition to the study that we have made in [1], which shows that users' current interests may help to get a better refined query, we propose that in some cases, users' social relations could help and social relation based interest models can help to refine the vague query too, since social relations serve as an environment for users when they perform query tasks.

From the perspective of scalable semantic search, this paper aims at achieving scalability through providing important search results to users. No matter how fast the data is growing for a semantic search system, the most important results for users will not grow dramatically. Users' social relation can be represented in the form of semantic data and serve as one kind of background information that can be used to help users acquire the most important search results.

In this paper, based on SwetoDBLP [2], an RDF version of the DBLP dataset, we provide some illustrative examples (mainly concentrating on expert finding and literature search) on how the social relations and social network based interest retention models can help to refine searching on the semantic data.

2 Social Relations and Social Networks

Social relations can be built based on friendship, coauthorship, work relationship, etc. The collection of social relationships of different users form a social network. As an illustrative example, we build a coauthor network based on the SwetoDBLP dataset, we represent the coauthor information for each author using FOAF vocabulary "foaf:knows".

The social network can be considered as a graph. In this example, each node is an author name and the relationships among nodes are coauthorships. An RDF dataset that contains all the coauthor information for each of the authors in the SwetoDBLP dataset has been created and released[5]. Through an analysis of node distribution for this DBLP coauthor network, we can find it has following statistical properties: As shown in Figure 1 and Figure 2 [3, 4]. The distribution can be approximately described as a power law distribution, which means that there are not many authors who have a lot of coauthors, and most of the authors are with very few coauthors. We can indicate that with this distribution characteristics, considering the scalability issue, when the number of authors expand rapidly, it will not hard to rebuild the coauthor network since most of the authors will just have a few links.

The purpose of this RDF dataset is not just to create a coauthor network, but to utilize this dataset to extract social relations from it and use them for refining the search process.

3 Search Refinement through Social Relationship

In enterprize information retrieval, expert finding is a emerging research topic [5]. The main task for this research area is to find relevant experts for a specific

[5] the coauthor network RDF dataset created based on the SwetoDBLP dataset can be acquired from http://www.iwici.org/dblp-sse

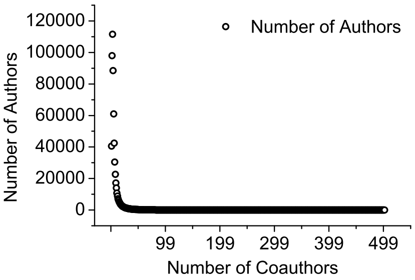

Fig. 1: Coauthor number distribution in the SwetoDBLP dataset.

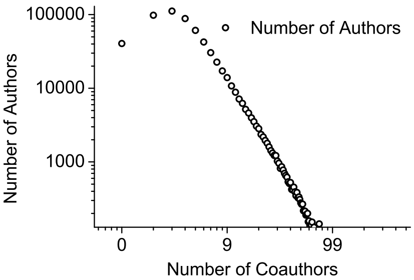

Fig. 2: log-log diagram of Figure 1.

domain [6]. Nevertheless, a list of expert names that has nothing to do with the end user always confuse them. More convenient search refinement strategies should be developed. We propose that if the end users are familiar with the retrieved expert names, the search results may be more convenient for use. As an illustrative example, we propose a search task that needs to find "Artificial Intelligence authors" based on the SwetoDBLP dataset.

Table 1: A partial result of the expert finding search task "Artificial Intelligence authors"(User name: John McCarthy).

Satisfied Authors without social relation refinement	Satisfied Authors with social relation refinement
Carl Kesselman (312)	Hans W. Guesgen (117) *
Thomas S. Huang (271)	Virginia Dignum (69) *
Edward A. Fox (269)	John McCarthy (65) *
Lei Wang (250)	Aaron Sloman (36) *
John Mylopoulos (245)	Carl Kesselman (312)
Ewa Deelman (237)	Thomas S. Huang (271)
...	...

Table 1 provides a partial result for the experiment of the proposed expert finding search task (Here we only consider a very simple and incomplete strategy, namely, find the author names who have at least one paper with "Artificial Intelligence" in its title). The left column is a partial list of results without social relation based refinement, which is just a list of author names without any relationship with the user. The right column is a partial list of results with social relation based refinement (The refinement is based on the social relations of the specified user that are extracted from the social network created in Section 2). Namely, the "Artificial Intelligence" authors whom the user "John McCarthy" knows are ranked to the front (As shown in the table, including himself). The results of the right column type seems more convenient for a user since the results which are ranked to the first ones seems to be familiar with the user compared to a list of irrelevant names. In a enterprise setting, if the found experts have some previous relationship with the employer, the cooperation may be smoother.

In this example, a user's collaborators appeared in two different scenarios, namely, in the coauthor network and domain experts knowledge base. Both of them are represented as semantic datasets using RDF, which enables the following connection. When a user tries to find domain experts, his social relations in the coauthor network are linked together with the domain experts knowledge base through the user's name or URI. This connection brings two separate datasets together and help to refine the expert finding task.

4 Social Network based Interest Retention Models for Search Refinement

In our previous study, we found that user interest retentions from users' historical information (e.g. previous publication list, web page log. etc.) can be acquired and used to refine the search process [1]. In addition, we propose that interest retentions of the users' friends, colleagues, collaborators may serve as environmental factors that affect users' query requirement (For example, may be most friends of a user are interested in a topic, and it motivates the user to investigate on this topic by him/herself). We first introduce the user interest models developed in [1], then we build a social network based interest model to track the "group interest" of a social network that a user is involved in. At last, we refine a search task based on the semantic dataset using the acquired group interest.

User Interests can be described as a set of concepts that users are interested in. For simplicity, we bring two interest models introduced in [1] to track a specific user's interests and interests of his/her collaborators.

Total interest(TI) function:

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

where $j \in [1, n]$ is a variable that is used to denote a certain time interval, and n is the number of time intervals (e.g. year) considered, $m(i, j)$ is the number of appearances of term i in the time interval j, $TI(i)$ reflects the value of total interest on topic i, namely, how many times has a interest appeared in the considered time intervals.

The above computation may not correctly reflect a researcher's current interests. For example, he/she has shifted the interest, but the accumulated number of an old interest may still be higher than that of a new interest. In [1], we emphasized that the interest retention, which is very related to a user's current interest, can be modeled by using memory retention like functions [7, 8]. Here we introduce one of them, the power law based model for interest retention calculation.

Interest retention function:

$$IR(i) = \sum_{j=1}^{n} m(i,j) \times AT_i^{-b}, \quad (2)$$

where T_i is the time interested in topic i until a specified year. For each time interval j, the interest i might appear $m(i,j)$ times, and $m(i,j) \times AT_i^{-b}$ is the total retention of an interest contributed by that time interval.

To sum up, $TI(i)$ reflects a user's interest on topic i through all the time intervals, which reflects the total interest value through all the considered time. $IR(i)$ reflects a user's current interest on topic i, and they focus on the interest retention on the topic in more recent years.

As an illustrative example, we consider a scenario of tracking the authors' research interests, which are implicitly embedded in their own publication lists. The time interval considered in here is a year. Since the interests retention might be related to users' current interests, we use the values from interest retention models to predict users' current research interests and by this way, we get the value for the parameter "A" and "b". As introduced in [1], in order to minimize the value of ρ in t-test, as a first try, the parameters in the power law based model are chosen as $A = 0.855$ and $b = 1.295$.

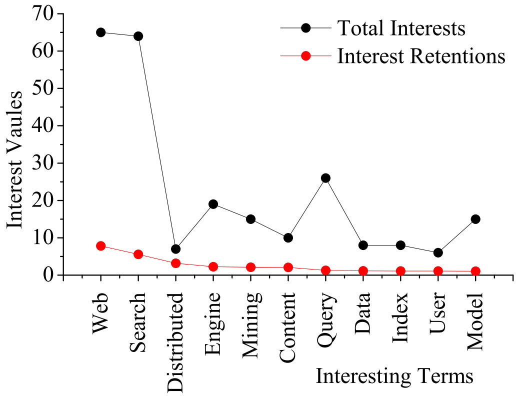

Fig. 3: A Comparison of Total Interests and Interest Retentions of the author "Ricardo A. Baeza-Yates".

Figure 3 shows a comparative study of total interests and interest retentions based on the data analysis of the author "Ricardo A. Baeza-Yates". As observed, an interest (i) with relatively high total interest value $(TI(i))$, does not always has a high interest retention vaule $(IR(i))$, such as "query" in the figure. In addition, although some of the interests, such as "distribution" does not have higher total interests, they may have very high interest retentions since they may be currently, at least most recently interesting to a user.

The following is a sample RDF file representing Ricardo A. Baeza-Yates's current research interests (using the power law based interest retention model) through analysis of his publications from the SwetoDBLP dataset:

```
<?xml version="1.0" encoding="ISO-8859-1"?> <!DOCTYPE rdf:RDF [
   <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#">
]> <rdf:RDF
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:rdfs = "http://www.w3.org/2000/01/rdf-schema#"
     xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <foaf:Person>
      <foaf:name>Ricardo A. Baeza-Yates</foaf:name>
      <rdfs:seeAlso rdf:resource=
      "http://www.informatik.uni-trier.de/~ley/db/indices/
      a-tree/b/Baeza=Yates:Ricardo_A=.html"/>
      <rdf:Seq>
      <foaf:topic_interest>Web</foaf:topic_interest>
      <foaf:topic_interest>Search</foaf:topic_interest>
      <foaf:topic_interest>Distributed</foaf:topic_interest>
      <foaf:topic_interest>Engine</foaf:topic_interest>
      ...
      </rdf:Seq>
  </foaf:Person>
</rdf:RDF>
```

where we use foaf:topic_interest to describe the author's interests and the RDF sequence container to describe the order of interests.

As a foundation for the development of social "group interest", we analyzed all the authors' interests retention values based on the SwetoDBLP dataset (more than 615,000 authors) using the introduced model, an RDF version of the interest enhanced DBLP author set has been released on the project page [6].

A user and his/her friends, collaborators form a social network. the user's interests may be affected by this social network since the network contains a group of other users who also have some interests. If they always communicate with each other, in the form of talking, collaboration, coauthoring, etc., their interests may be affected by each others'. If the user is affected by "group interests", he/she may begin to search on the interesting topic to find relevant information. Hence, group interests may be serve as an essential environmental factor from user background for search refinement. The group interests can be acquired through linking the user interests semantic dataset and the social network semantic dataset (They can be linked together by user names or URIs).

Social Network based Group Interest Retention functions:

$$GIR(i) = \sum_{k=1}^{p} IR_k(i) = \sum_{k=1}^{p} \sum_{j=1}^{n} m_k(i,j) \times AT_{i,k}^{-b}, \tag{3}$$

[6] http://www.iwici.org/dblp-sse and http://wiki.larkc.eu/csri-rdf

where p is the number of collaborators (or friends, etc.) who are directly connected to a specified user, $GIR(i)$ is the value of the group interest retention for an interest i, $T_{i,k}$ is the time interested in the topic i by the collaborator (or friend, etc.) k until a specified year. For each time interval j, the interest i might appear $m_k(i, j)$ times for the collaborator (or friend, etc.) k, and $m_k(i, j) \times AT_{i,k}^{-b}$ is the total retention of an interest for the collaborator k contributed by that time interval.

Another simplified alternative is to count the appear time of the top N retained interests in the group of collaborators, and see which interests can be ranked to the front as group interest retentions.

$$GIR(i) = \sum_{k=1}^{p} E(i), \tag{4}$$

where $E(i) \in \{0, 1\}$, if the interest i appears in the top N retained interests of a user, then $E(i) = 1$, otherwise, $E(i) = 0$.

Using formula (2) and (4), take "Ricardo A. Baeza-Yates" as an example, a comparative list of top 7 interests retention of his own and his group interests retention (with 132 authors involved) is shown in Table 2.

Table 2: Top 7 interests retention of a user and his group interests retention. (User name: Ricardo A. Baeza-Yates)

Web	7.81	Search (*)	35
Search	5.59	Retrieval	30
Distributed	3.19	Web (*)	28
Engine	2.27	Information	26
Mining	2.14	System	19
Content	2.10	Query (*)	18
Query	1.26	Analysis	14

Through Table 2 we can find that group interest retentions are not the same as, but to some extent related to the user's own retained interests (interesting terms that are marked with "*" are the same). Hence, group interest retentions and user's own retained interests can be used as two sources to refine the search process and satisfy various user needs.

Table 3 shows a comparative study of search results without refinement, with user retained interests based refinement, and with group retained interests based refinement. Different search results are selected out and provided to users to meet their diverse needs. One can see that how the group interests serve as an environmental factor that affect the search refinement process and help to get more relevant search results.

Table 3: Search Refinement using user's own retained interests and group interest retentions

Name: Ricardo A. Baeza-Yates
Query : Intelligence
List 1 : without any refinement (top 7 results)
1. PROLOG Programming for Artificial **Intelligence**, Second Edition. 2. Artificial **Intelligence** Architectures for Composition and Performance Environment. 3. The Mechanization of **Intelligence** and the Human Aspects of Music. 4. Artificial **Intelligence** in Music Education: A Critical Review. 5. Readings in Music and Artificial **Intelligence**. 6. Music, **Intelligence** and Artificiality. 7. Regarding Music, Machines, **Intelligence** and the Brain: An Introduction to Music and AI.
List 2 : with user's own interests constraints (top 7 results)
interests : Web, Search, Distributed, Engine, Mining, Content, Query
1. SWAMI: Searching the **Web** Using Agents with Mobility and **Intelligence**. 2. Moving Target **Search** with **Intelligence**. 3. Teaching **Distributed** Artificial **Intelligence** with RoboRally. 4. Prototyping a Simple Layered Artificial **Intelligence Engine** for Computer Games. 5. Web Data **Mining** for Predictive **Intelligence**. 6. **Content** Analysis for Proactive **Intelligence**: Marshaling Frame Evidence. 7. Efficient XML-to-SQL **Query** Translation: Where to Add the **Intelligence**?
List 3 : with group retained interests constraints (top 7 results)
interests : Search, Retrieval, Web, Information, System, Query, Analysis
1. Moving Target **Search** with **Intelligence**. 2. A New Swarm **Intelligence** Coordination Model Inspired by Collective Prey **Retrieval** and Its Application to Image Alignment. 3. SWAMI: Searching the **Web** Using Agents with Mobility and **Intelligence**. 4. Building an **information** on demand enterprise that integrates both operational and strategic business **intelligence**. 5. An Explainable Artificial **Intelligence System** for Small-unit Tactical Behavior. 6. Efficient XML-to-SQL **Query** Translation: Where to Add the **Intelligence**? 7. **Intelligence Analysis** through Text Mining.

5 Evaluation and Analysis

Since the user interests and group interests are obtained from analysis based on real authors in the DBLP system. For the evaluation of the experimental results, the participants also need to be real authors in the system, preferably those with some publications distributed in different years. These constraints made finding the participants not easy.

Currently, we have received evaluation results from 7 authors that have some publication listed in DBLP. Through an analysis of these results, we find that: 100% of these authors feel that the refined search results using user interests retention and group interests retention are much better than the result list which does not have any refinement. 100% of them feel that the satisfaction degree of the two refined result lists are very close. 83.3% of them feel that refined results by the users' own interests retention is better than others. 16.7% of them feel refined results by group interest retentions are better than others.

The refined list with the authors' own recent interests retention is supposed to be the best one. Since the query constraints are all most related information that the users are interested in. We need to explain why the group interests retention could also help to get much better search results. If we add all of the coauthors' interests retention together for an author, we observe that most of the biggest interests are quite relevant to his/her own one. We randomly pick 30 authors from the SwetoDBLP dataset and we calculate their own interests retention and group interests retention. We find that in average, the biggest 8 interests retention and group interests retention have 57% interests in common. That's why the refined list with group interests retention are also welcome and considered much better than the one without any refinement. Interests from the author's social network are very relevant to his/her own interests! It indicates that if one's own interests can not be acquired, his/her friends' interests also could implicitly help to refine the search process and results.

6 Conclusion and Future Work

In this study, we provide some illustration on how the social relations and social network based interest retention models can help to refine searching on large scale semantic data. For the scalability issue, this approach scales in the following way: No matter how large the semantic dataset is, through the social relation based interests retention model, the amount of most relevant results are always relatively small, and they are always ranked to the top ones for user investigation.

The methods introduced in this paper are related but different from traditional collaborative filtering methods [9, 10]. Firstly, both the user and their friends (e.g. coauthors, collaborators) do not comment or evaluate any search results (items) in advance. Secondly, interest retention models (both users' own one and their group one) track the retained interests as time passed. The retained interests are dynamically changing but some of the previous interests are retained according to the proposed retention function. Thirdly, as shown in Section 3, user interests stored in different data sources are linked together for search refinement. There is no need to have social relation information in one dataset or system. Another example is that, if someone in the DBLP system wants to buy books on Amazon, he/she does not have to have a social relation on Amazon which can be used to refine the product search. Through the linked data from the group retained interests based on SwetoDBLP, the search process also could be refined.

For now, semantic similarities of all the extracted terms have not been added into the interest retention models. Some preliminary experiments show that this may reduce the correlation between an author's own interests retention and his/her group interests retention. For example, for the user "Guilin Qi", both his current retained interests and his group interests contain "OWL" and "ontology", which seem to be 2 different terms. But in practice, "OWL" is very related to “Ontology” (for their Normalized Google Distance [11], $NGD(ontology, owl) = 0.234757$, if $NGD(x, y) \leq 0.3$, then x and y is considered to be semantically very related [11]). For the user “Zhisheng Huang”, the terms “reasoning” and “logic” are 2 important interests, while reasoning is very related to “logic” ($NGD(logic, reasoning) = 0.2808$). In our future work, we would like to use Google distance [11] to calculate the semantic similarities of interesting terms so that more accurate retained interests can be acquired and better search constraints can be found. We also would like to see whether other social network theories (such as six degree of separation) could help semantic search refinement in a scalable environment.

7 Acknowledgement

This study is supported by the research grant from the European Union 7th framework project FP7-215535 Large-Scale Integrating Project LarKC (Large Knowledge Collider). We thank Yang Gao for his involvement on the program development of interest retentions for authors in the SwetoDBLP dataset.

References

1. Zeng, Y., Yao, Y., Zhong, N.: Dblp-sse: A dblp search support engine. In: Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence. Volume 1., IEEE Computer Society (September 2009) 626–630
2. Aleman-Meza, B. Hakimpour, F., Arpinar, I., Sheth, A.: Swetodblp ontology of computer science publications. Web Semantics: Science, Services and Agents on the World Wide Web **5**(3) (2007) 151–155
3. Elmacioglu, E., Lee, D.: On six degrees of separation in dblp-db and more. SIGMOD Record **34**(2) (2005) 33–40
4. Zeng, Y., Wang, Y., Huang, Z., Zhong, N.: Unifying web-scale search and reasoning from the viewpoint of granularity. In: Proceedings of the 2009 International Conference on Active Media Technology, Springer (October 2009) 418–429
5. Balog, K., Azzopardi, L., de Rijke, M.: Formal models for expert finding in enterprise corpora. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (2006)
6. YimamSeid, D., Kobsa, A.: ExpertFinding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. In: Sharing Expertise: Beyond Knowledge Management. 1 edn. The MIT Press (2003) 327–358
7. Wickelgren, W.: Memory storage dynamics. In: Handbook of learning and cognitive processes. Hillsdale, NJ: Lawrence Erlbaum Associates (1976) 321–361
8. Anderson, J., Schooler, L.: Reflections of the environment in memory. Psychological Science **2**(6) (1991) 396–408
9. Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Communications of the ACM **35**(12) (1992) 61–70
10. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: An open architecture for collaborative filtering of netnews. In: Proceedings of the Conference on Computer Supported Cooperative Work. (1994) 175–186
11. Cilibrasi, R., Vitanyi, P.M.B.: The google similarity distance. IEEE Transaction on Knowledge and Data Engineering **19**(3) (2007) 370–383