Social Relation Based Search Refinement : Let Your Friends Help You!

Xu Ren[1], Yi Zeng[1], Yulin Qin[1,2], Ning Zhong[1,3],
Zhisheng Huang[4], Yan Wang[1], Cong 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 search at Web scale is that the search results on the large scale 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 although the input query may be the same. In this paper, we try to achieve scalability for Web search through social relation diversity of different users. Namely, we utilize one of the major context for users, social relations, to help refining the search process. Social network based group interest models are developed according to collaborative networks, and is designed to be used in more wider range of Web scale 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 the Web environment.

Keywords. social relation, retained interest, social network based group interest model, personalized 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 goes 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 users are interested in or at least familiar with. In addition to the study that we have made in [1], which shows that users' recent interests may help to get a better refined query, we propose that in some cases, users' social relations and social network based group 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 Web search, this paper aims at achieving scalability through providing important search results to users. Since no matter how fast the data is growing, the size of the most important search results for users will be relatively small. 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 group interest models can help to refine searching on the Web.

2 Social Relations and Social Networks

Social relations can be built based on friendships, coauthorships, work relationships, 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. Each node can be an author name and the relationships among nodes can be 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 be 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.

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

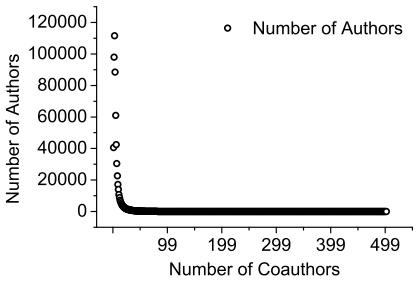

Fig. 1: Coauthor number distribution in the SwetoDBLP dataset.

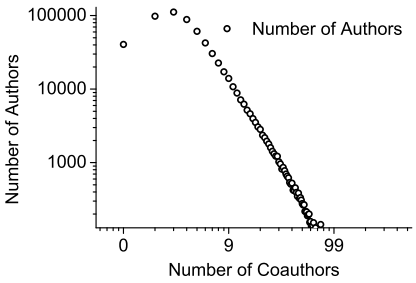

Fig. 2: log-log diagram of Figure 1.

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 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 (here we consider SwetoDBLP as the 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 Group Interest Models

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 the social network based "group interests", he/she may begin to search on the interesting topic to find relevant information. Hence, the social network based group interests may be serve as an essential environmental factor from user background for search refinement.

Group Interest:

For a specific interest "t(i)", its group interests for a specific author "u", namely "$GI(t(i), u)$" can be quantitatively defined as:

$$\begin{aligned} GI(t(i), u) &= \sum_{c=1}^{m} E(t(i), u, c), \\ E(t(i), u, c) &= \begin{cases} 1 \ (t(i) \in I_c^{topN}) \\ 0 \ (t(i) \in I_c^{topN}) \end{cases} \end{aligned} \tag{1}$$

where $E(t(i), u, c) \in \{0, 1\}$, if the interest $t(i)$ appears both in the top N interests of a user and one of his/her friends', then $E(t(i), u, c) = 1$, otherwise, $E(t(i), u, c) = 0$. For a specific user "u", there are m friends in all, and the group interest of "t(i)" is the cumulative value of $E(t(i), u, c)$ based on the m friends. In a word, group interest focuses on the cumulation of ranked interests from a specific user's social network.

Various models can be used to quantitatively measure and rank interests so that one can get the top N interests to produce the value of group interests. We

defined 4 models in [7], here we briefly review them so that the comparison of group interests from the 4 perspectives can be made.

Let i, j be positive integers, $y_{t(i),j}$ be the number of publications which are related to topic $t(i)$ during the time interval j.

Cumulative Interest:

Cumulative interest, denoted as $CI(t(i), n)$, is used to count the cumulative appear times of the interest $t(i)$ during the n time intervals. It and can be acquired through:

$$CI(t(i), n) = \sum_{j=1}^{n} y_{t(i),j}. \tag{2}$$

It is used reflect a user's over all interest on the specified topic within a time interval.

Retained interest:

A person may be interested in a topic for a period of time but is likely to loose interest on it as time passes by if it has not appeared in some way for a long time. This phenomenon is very similar to the forgetting mechanism for cognitive memory retention. In [1] we introduced an retained interest model based on a power law function that cognitive memory retention [8] follows:

$$RI(t(i), n) = \sum_{j=1}^{n} y_{t(i),j} \times AT_{t(i)}^{-b}, \tag{3}$$

where $T_{t(i)}$ is the duration interested in topic $t(i)$ until a specified time. For each time interval j, the interest $t(i)$ might appear $y_{t(i),j}$ times, and $y_{t(i),j} \times AT_{t(i)}^{-b}$ is the total retention of an interest contributed by that time interval. According to our previous studies, the parameters satisfy $A = 0.855$ and $b = 1.295$ [1].

Interest Cumulative Duration:

Interest cumulative duration, denoted as $ILD(t(i))$, is used to represent the longest duration of the interest $t(i)$:

$$ILD(t(i)) = \max(ID(t(i))_n). \tag{4}$$

where $n \in I^+$, $ID(t(i))_n$ is the interest duration when $t(i)$ discretely appears (the time interval of the appeared interest is not directly continuous with the one of the previous appeared interest) for the nth time.

Interest Longest Duration:

Interest longest duration, denoted as $ICD(t(i))$, is used to represent the cumulative duration of the interest $t(i)$:

$$ICD(t(i)) = \sum_{n=1}^{n'} (ID(t(i))_n). \tag{5}$$

where $n \in I^+$ is used to represent the nth discrete appearance of the interest $t(i)$, and n' is the total discrete appearance times of the interest $t(i)$.

The above 4 interest models are used for producing the top N interests. The corresponding group interests based the proposed models are: group cumulative interest ($GCI(t(i), u)$), group retained interest ($GRI(t(i), u)$), group cumulative duration ($GCD(t(i), u)$), and group longest duration ($GLD(t(i), u)$) respectively. Their calculation function is the same as $GI(t(i), u)$, namely, $GCI(t(i), u)$, $GRI(t(i), u)$, $GCD(t(i), u)$ and $GLD(t(i), u)$ are special cases of $GI(t(i), u)$.

As a foundation for the development of social "group interest", we analyzed all the authors' retained interests 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].

Here we give an illustrative example on producing group interests based on retained interests. Using formula 3 and 1, and taking "Ricardo A. Baeza-Yates" as an example, a comparative list of top 7 retained interests of his own and his group retained interests (with 132 authors involved) is shown in Table 2.

Table 2: A comparative study of top 7 retained interests of a user and his/her group retained interests. (User name: Ricardo A. Baeza-Yates)

Self Retained Interests	Value	Group Retained Interests	Value
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 may be group retained interests 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).

As a step forward, we analyzed the overlap between a specific user's own interests and his/her group interests. 50 most productive authors from the DBLP dataset (May 2010 version) are selected for the evaluation. The analysis considers 4 types of overlaps:

- cumulative interest ($CI(t(i), n)$) and group cumulative interest ($GCI(t(i), u)$),
- retained interest ($RI(t(i), n)$) and group retained interest ($GRI(t(i), u)$),
- interest longest duration ($ILD(t(i), j)$) and group interest longest duration ($GLD(t(i), u)$),
- interest cumulative duration ($ICD(t(i), j)$) and group interest cumulative duration ($GCD(t(i), u)$).

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

The value of the overlaps are average values of the selected 50 authors. As shown in Figure 3, from the 4 perspectives, the overlaps range are within the interval [0.593, 0.667]. It means that no matter from which of these perspectives, the overlap between the users' own interests and their group interests are at least greater than 59%. Take $RI(t(i), n)$ vs $GRI(t(i), u)$ and $CI(t(i), n)$ vs $GCI(t(i), u)$ as examples, Figure 4 shows that for most of the 50 authors, the overlaps are within the time interval [0.4,0.9].

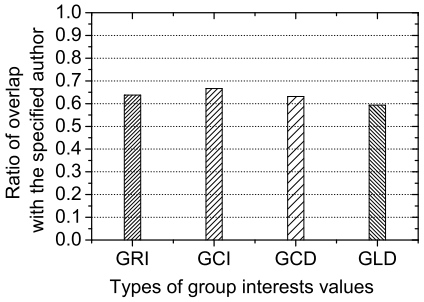

Fig. 3: Ratio of Overlap between different group interest values and the specified author's interest values

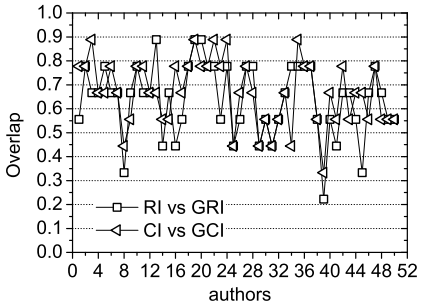

Fig. 4: A comparative study on the overlap between RI and GRI, CI and GCI.

Based on the results and analysis above, besides users' own interests, their group interests and can be used as another source to refine the search process and satisfy various user needs.

5 Group Interests Based Search Refinement

In [1], according to the idea of retained interests model of a specific user, we developed a DBLP Search Support Engine (DBLP-SSE), which utilizes the user's own retained interests to refine search on the SwetoDBLP dataset [2]. Based on the idea of group retained interest model introduced in 4, we developed a search support engine based on the SwetoDBLP dataset [2]. Figure 5 is a screen shot on the current version of the DBLP Search Support Engine (DBLP-SSE).

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 social network based group interests serve as an environmental factor that affect the search refinement process and help to get more relevant search results.

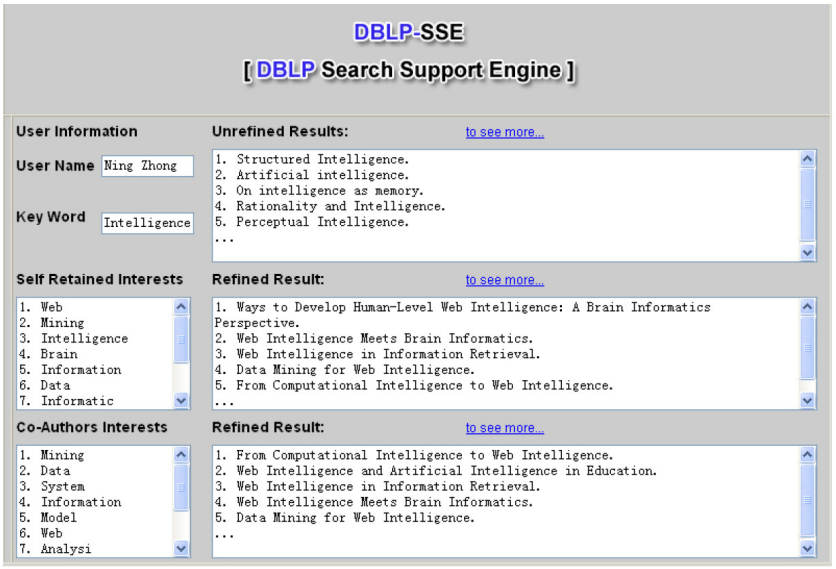

Fig. 5: A screen shot of the DBLP Search Support Engine (DBLP-SSE).

6 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.

The participants are required to search for intelligence in the DBLP Search Support Engine (DBLP-SSE)[7] that we developed based on the SwetoDBLP dataset [2]. Three lists of query results are provided to each of them. One is acquired based on unrefined query, and another two are refined by users' own top 9 retained interests and top 9 group retained interests. They are required to judge which list of results they prefer.

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 users' most recent $RI(t(i), n)$ and $GRI(t(i), u)$ 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 $RI(t(i), n)$ is better than others. 16.7% of them feel refined results by $GRI(t(i), u)$ are better than others.

[7] DBLP-SSE is available at http://www.wici-lab.org/wici/dblp-sse

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

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.

The refined list with the authors' $RI(t(i), n)$ is supposed to be the best one. Since the query constraints are all most related information that the users are interested in. Since the average overlap between users' $RI(t(i), n)$ and $GRI(t(i), u)$ is around 63.8%, which means that interests from the author's social network are very relevant to his/her own interests! That's why the refined list with $GRI(t(i), u)$ are also welcome and considered much better than the one without any refinement. It indicates that if one's own interests can not be acquired, his/her friends' interests also could help to refine the search process and results.

7 Conclusion and Future Work

In this study, we provide some illustration on how the social relations and social network based interest models can help to refine searching on large scale data.

For the scalability issue, this approach scales in the following way: No matter how large the dataset is, through the social relation based group interest models, 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, following the idea of linked data [11], there is no need to have relevant information in one dataset or system. As shown in Section 3, user interests stored in different data sources are linked together for search refinement (user interests data and collaboration network data). Another example is that, if someone is recorded 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 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 retained interest models. Some preliminary experiments show that this may reduce the correlation between an author's own retained interests 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 [12], $NGD(ontology, owl) = 0.234757$, if $NGD(x, y) \leq 0.3$, then x and y is considered to be semantically very related [12]). 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 [12] 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.

8 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 Yiyu Yao for his idea and discussion on Search Support Engine, 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. (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. Volume 5820 of Lecture Notes in Computer Science., 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. Zeng, Y., Zhou, E., Qin, Y., Zhong, N.: Research interests : Their dynamics, structures and applications in web search refinement. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligenc. (2010)
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. Bizer, C.: The emerging web of linked data. IEEE Intelligent Systems **24**(5) (2009) 87–92
12. Cilibrasi, R., Vitanyi, P.M.B.: The google similarity distance. IEEE Transaction on Knowledge and Data Engineering **19**(3) (2007) 370–383