AuthorRank: Ranking Improvement for the Web

Semantic Web and Web Services Conference
Keywords: Social Networking, Page Rank, Semantic Web

Ying Ding and François Scharffe
Digital Enterprise Research Institute,
University of Innsbruck, Austria
Email: firstname.lastname@deri.org
Andreas Harth and Adrian Hogan
Digital Enterprise Research Institute,
University of Galway, Ireland
Email: firstname.lastname@deri.org

***Abstract*— As the wealth of data on the World Wide Web grows, and as the structuring of that data improves, more sophisticated applications can be developed to derive meaningful characteristics relating to the content and structure of that data. In particular, ranking the various elements of sets of structured information is of great utility with respect to semantic network analysis. In this paper we report on preliminary results of ranking experiments carried out on the DBLP dataset that contains metadata descriptions of more than 600.000 publications.**

I. Introduction

Google generated the revolution for the web search based on important ranking algorithm - PageRank which manages to bring the most relevant search results to the top of the returned result [1]. PageRank assumes the web hyperlinks as the trust votes and ranks the search results based on them. PageRank creates the new synergy to information retrieval for the better ranking of the Web. Researches migrate from traditional ranking algorithm based on keywords to various ranking algorithm to further improve PageRank. Since Google goes to commercial, it is hard to know how Google is ranking the results now. But PageRank is still claimed by Google as the key algorithm for ranking[1].

PageRank is not a new concept in information retrieval. Actually there is long history of citation research originated from 1940s. Before the web appears, printed journals, magazines or conference proceedings are the main publication channels of academic scholars. Fortunately these printed materials have the controlled format to follow where the quality of citations can be guaranteed at certain degree. Citation analysis, especially co-citation analysis, constructs an innovative way to analyze and rank documents. The hidden information and relations of the concepts can be mined by the co-citation analysis (such as co-word, co-author, co-journal, etc.). The result of co-citation analysis has been used for query extension [2], field analysis [3], visualizing intellectual structure of the field [4] and so on. This can be viewed as the early effort of social network analysis based on academic citations.

World Wide Web has accumulated tremendous amount of data and information which brings social network analysis into the new era. Nowadays researchers do not face the problem of the lack of data, on the contrary, they have to solve the problem of the overloaded data and the quality of the data. Methods and algorithms which work perfectly before might fail completely on the Web. Web as the rich repository creates the new challenge for scalability and efficiency. Computing huge co-occurrence matrix for the Web becomes very inefficient and nearly impossible. Finding efficient algorithm to handle large size of matrix for clustering and scaling is challenging, research on large graph study like the work of [5] on small-world networks is in that perspective interesting.

Semantic Web, as the next generation of the Web, produces meaningful data to the Web. On the one hand, it increases the quality of the Web data. On the other hand, it enriches the semantics of the data by adding metadata and ontologies. FOAF is one of the Semantic Web efforts. It brings more possibilities to rank the web information, such as ranking people based on FOAF[2] data - here is called AuthorRank. This paper aims to exploit AuthorRank based on FOAF and DBLP data and

[1] http://www.google.com/technology/

[2] http://xmlns.com/foaf/0.1/

the various combination of AuthorRank with cocitation analysis (such as co-author and co-word) in order to identify the efficient ranking algorithm for the web search, especially targeted for people search.

PeopleRank research has been conducted in Stanford Digital Library Project with quite different focus. [6] finds efficient algorithm to identify people in the photo albums based on context and labels of the photos. Based on the background knowledge of the authors and the extensive search of the Web (Google search), we could not find previous work on AuthorRank based on FOAF data and the combination of the AuthorRank with cocitation analysis. Therefore we deem our research quite innovative. Specifically, we make the following contributions:

1) we show how to interrelate different datasets from the Web and combine rankings of different data sets into a combined ranking
2) we exploit social relations between people to propagate topics of interest for clustering of papers into topics

This paper contains the following sections. Section 2 explains how we collected FOAF data from the Web and how we extracted data from DBLP. Section 3 shows the AuthorRank algorithm and discusses the test results based on AuthorRank and the combination of AuthorRank with co-word and co-author analysis. Section 4 mentions the future work.

II. Dataset

In the following section we describe the dataset we collected from the Web. The data is stored and queried using YARS [7], a scalable RDF storage and querying system.

A. Friend-of-a-Friend (FOAF)

The Friend of a Friend (FOAF) project is about creating a Web of machine-readable homepages describing people, the links between them and the things they create and do[3]. FOAF is one of the most widely used vocabularies on the Semantic Web to date. FOAF is used in many scientific works, like in [8]. Figure 1 shows an example of a FOAF file describing two of the authors. The properties we utilitized in our experiments are foaf:name (which denotes the name of a person) and foaf:knows (which denotes that person A knows person B).

[3]http://www.foaf-project.org/

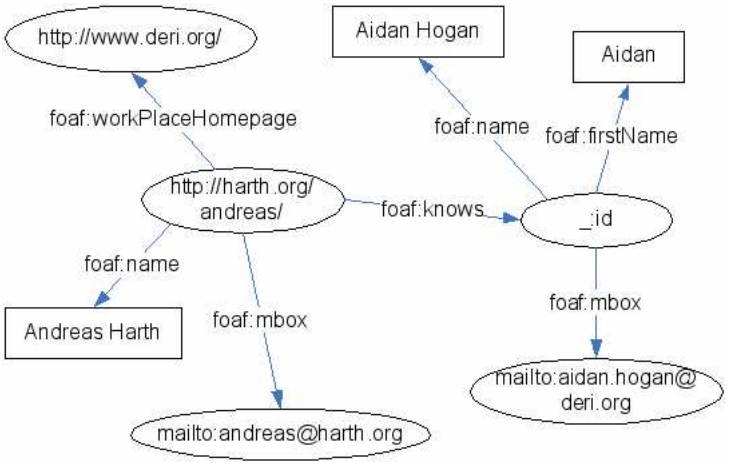

Fig. 1. Description of two of the authors in FOAF vocabulary

FOAF data is usually published in flat files on people's homepages and linked via rdfs:seeAlso predicates. We used a modified version of AIFB's RDF crawler[4] to retrieve FOAF files from the Web. This RDF crawler downloads interconnected fragments of RDF from the Internet and builds a knowledge base from these data. At each phase of RDF crawling, a list of URIs to be retrieved as well as URI filtering conditions (eg. Depth, URI syntax) should be provided. This RDF crawler is a stand alone application, which is given URIs and builds an RDF database from it or extends existing database. The dataset we used in our experiments[5] contained 34709 foaf:name relations and 22175 foaf:knows relations.

B. DBLP

The computer science bibliography dataset DBLP[6] contains descriptions about more than 600.000 publications in the area of Computer Science. To be able to combine DBLP data with the FOAF dataset we obtained from the web, we converted the publicly available XML version of DBLP[7] (2005-07-04 version) into RDF/XML[8]. The total size of the dataset in RDF/XML is 435 MB. Figure 2 shows two publications from DBLP; the papers are connected to each other by common authors.

Although there is citation data available, there are only few papers which include citation links, which is

[4]http://ontobroker.semanticweb.org/rdfcrawl/
[5]http://sw.deri.org/2005/04/semwebbase/
[6]http://dblp.uni-trier.de/
[7]http://dblp.uni-trier.de/xml/
[8]http://sw.deri.org/ aharth/2004/07/dblp/

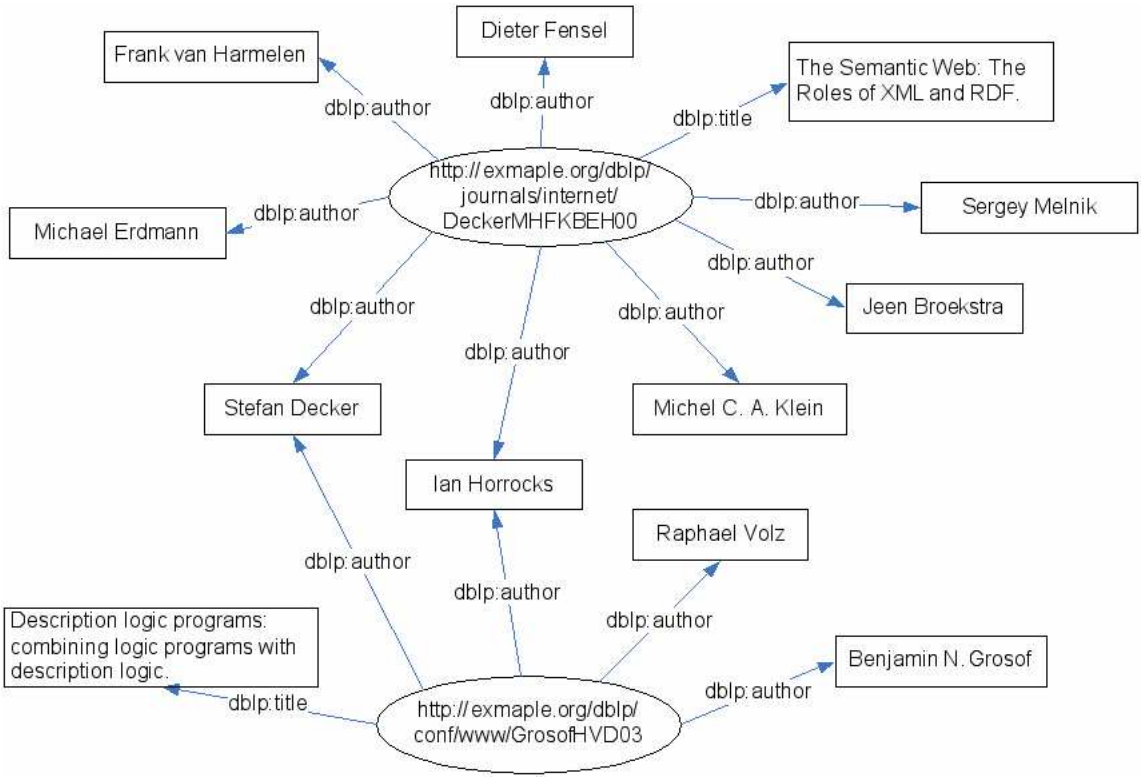

Fig. 2. Two papers from DBLP that are related to each other via common authors.

a restriction in the way we can compute a rank for the authors; traditional citation analysis doesn't work due to the incompleteness of citation data. But the current joint effort between the ACM SIGMOD Anothology and DBLP provides potential for future citation analysis. They try to provide citation links for an increasing number of publications in the area of database systems. Up till now, more than 100000 links have been entered. The citation information includes not only the traditional 'reference' section but also the 'referenced by' section[9]. This clearly points out the future direction of the current work on the combination of citation analysis with the ranking algorithm based on FOAF. More details are mentioned in section 4.

III. Experiments

In the following we describe the experiments we have carried out, together with results. We calculate a global rank for each author based on the dblp:author relation (co-author) and the foaf:knows relation. We then cluster the publications into subjects based on bigram analysis (bigram on word - the name of the author) and topic propagation along the dblp:author and foaf:knows relations. Finally, we combine the two ranking methods to generate an author rank in a given field. The underlying assumption is that people work on similar subjects. We propagate topics along the co-author relation to spread the topic bigram even if the bigram itself doesn't occur in the title.

A. Representing the link structure of the graph

Within the DBLP dataset, there are over 400,000 unique authors. A graph exists within this dataset where people are linked by co-authoring with one another. The most common method of representing the link structure of such a graph would be with a square connectivity matrix containing a 1 at index (i, j) if a link exists from node i to node j and a 0 otherwise. However, the memory requirements of such a square matrix would be vast. The nature of the graph would dictate that the matrix would be highly sparse. In order to make the ranking of such a graph efficient, and in order to create a technique which would be highly scalable, a scheme was devised to denote the link structure. Initially, one 'flat' matrix was used to represent the data. As with the customary square matrix, each node in the graph has a row of the matrix assigned to it. The identifier of each node, in this case

[9]http://www.informatik.uni-trier.de/~ley/db/conf/sigmod/CareyDN93.html

the full name of the author, is mapped to a unique row index. The conventional method would involve having a column for each node also, but this is highly expensive with regards to memory. Instead with the flat matrix, in the row of the relevant node, the indices of the nodes to which it links (out-links) are stored, so the average row length is greatly diminished for a sparsely linked graph. In a graph with n nodes and m links, the number of elements needed to construct a conventional connectivity matrix would be $n * n$. The number of elements in the flat matrix equivalent would be m. The flat matrix still holds the correspondent data but is significantly smaller. Unfortunately, it was found that this representation alone was quite slow when a lookup of the in-links of a particular node was required. To counteract this problem, a second matrix was introduced, the second matrix being of similar composition to the first, but it's inverse. Along the row belonging to a particular node are positioned the indices of nodes with in-links to that node. With these two matrices, the size is $2 * m$ and it is as efficient as the conventional connectivity matrix with concern to lookups. In addition to this, the nature of the ranking desired would require weightings to improve results. An extra dimension was added to each of the matrices to contain this data. The weight for a link found in a particular index of the former matrix was stored in the same index in the new dimension.

B. Link Based Analysis

For the purpose of ranking authors from the co-authors graph, the algorithm used should have a predilection towards authors who have collaborated with other important authors on numerous papers. The authority flow nature of the prevalent PageRank algorithm would seem to fit this mould. The PageRank technique also fits some of the other requirements raised by this use case, in that it is superbly scalable and highly flexible.

In experiments involving usage of such a system of ranking, some interesting issues arose. Within the co-author dataset there exists various sub-graphs, the most common being authors who have not co-authored with anyone. Also these was a rather large sub-cluster comprising of about 20 authors who co-authored with one another but not with any of the other authors from the main graph. Results from the earlier nave versions of the analysis were heavily skewed towards such sub-clusters, with people who had not co-authored with anyone or members of any sort of sub-graph receiving ratings beyond their merit. In fact, in the first experiment with the earliest version of the technique, the most prolific of the authors within the large sub-graph of 20 authors was rated number 1.

After investigation of the possible reasons, it was ascertained that self-links should be omitted from the link data. In earlier versions, it was considered that links existed from a node to itself, mainly stemming from the fashion in which the raw data set was harvested. This was causing severe accumulation of authority in sub-graphs. After removal of self-links, results improved considerably, with sole authors being relegated to the very bottom of the ranking list and member nodes of sub-clusters descending numerous positions. However, scope for augmenting the algorithm was still evident, this scope taking the form of weightings.

Previously, a person who had co-written a paper with an important author once, and a person who had collaborated with important author frequently, would receive the same authority flow from that author. Intuitively, if rank weightings were introduced to reflect the number of times one co-authored with another, results would be enhanced. In our final experiment analyzing the co-author matrix, a count of the number of links present between each of the nodes was maintained and used as weightings for the algorithm.

Five sets of results now exist for the 2 DBLP graphs, co-occurrence and co-author. Of particular interest should be the co-author results. In all results sets, the most highly ranked are at the bottom. All the results are derived from a ranking of the whole graph, the whole co-occurrence and the whole co-author graph. The ranking algorithms used are variations of a Page Rank algorithm, which would seem to be most beneficial to the co-author graph, where Page Rank is an authority flow algorithm. Co-authoring with important authors would reflect in a higher ranking than co-authoring with lower ranked authors.

- In the first results set[10], self links were included, and weights excluded. This resulted in undesirable results however as subgraphs received abnormally high ranks. People who just authored papers on their own were receiving incremental increases in their ranking brought on by the weak global link, the damping factor in the algorithm. These people were inherent subgraphs within the data set and were getting far more ranking than they deserved. The highest rank author was himself participating in a rather large subgraph. Of all the people he authored papers with, none of them authored with anyone

[10]http://sw.deri.org/ãidanh/PplRank/CArank0.n3

from the main graph, he was authoring within a secluded group of authors, and so the global link was incrementing their ranking abnormally. Common sense would deem that he was not deserving of the top spot.

- In the second set of results[11] self-links were excluded this time, and weights were still excluded. This completely solved the issue of people authoring papers on their own, forming their own mini subgraph. All these people were now the lowest ranked. Also, the top spot was filled by a more deserving author. It did not completely solve the issue though, as the author who was at the top in the former results set, was still an undeserving fifth. Not only did removing these self-links improve results, but the algorithm converged to an acceptable tolerance after only less than half the iterations, 17.
- In the third set[12], the number of times a person co-authored with another was taken into consideration, weighted links were used. This was a step beyond the usual Page Rank algorithm, which would hopefully boost the ratings of authors who collaborated numerous times with other important authors. In this set, it would seem, those more highly ranked are much more deserving of their place.

C. Clustering of Papers

To be able to construct not only a global rank of authors, but a rank of authors within a given field, we perform clustering of papers. Our approach is a combination of bigram construction with weight propagation. We next show in pseudocode the steps we used to construct topic descriptions based on bigrams.

```
1. get list of bigrams of the words
in the dataset and the occurrence
frequency of the bigrams
(only words of length 3 or greater
are considered)
2. sort the list, eliminate bigrams
that contain stopwords
('and', 'the', 'der', 'die', ...)
We use a list of stopwords.
3. the first 1000 bigrams serve as topic
description;
each bigram is converted into a URI,
and each paper with the bigram in its
title gets assigned the topic URI
```

[11]http://sw.deri.org/ãidanh/PplRank/CArank1.n3

[12]http://sw.deri.org/ãidanh/PplRank/CArank1.n3

Next is the algorithm to propagate the topic descriptions along the author relation.

```
1. get all papers of authors
that have written a 'Semantic Web'
paper
2. construct a matrix where the
'Semantic Web' papers get a weight
of '1' and the non-semantic web papers
get a weight of '0'.
3. run a couple of iterations
(why not until it reaches an equilibrium?)
and propagate the rank
(running time on a Intel Pentium-4 laptop:
 couple of seconds)
```

The result of the calculation is that each paper gets assigned a rank, which specifies the probability that the given paper is part of the topic 'Semantic Web'. Table I shows the top 20 papers that were assigned the topic 'Semantic Web' by our algorithm without including the bigram 'Semantic Web' in the title. The papers in the top 20 list are papers with many authors.

D. Combining Clusters with Global Rank

We have described how to calculate global ranking scores based on DBLP co-author relations, and foaf:knows relations. Also, we have shown how to cluster documents and authors around topic descriptions. In the following, we combine the results from both global rank calculation and clustering to be able to generate ranks for the most important authors that have published papers in the Semantic Web field in the FOAF sphere (Table II) and the DBLP-sphere (Table III).

IV. Future Work

This research aims to identify an efficient ranking algorithm for people search on the Web. Although it is still at the early stage, some interesting results have been presented here. Based on the literature review of related works, our research on AuthorRank and the combination of AuthorRank with co-word and co-author analysis is innovative and has not been done before. Since DBLP puts effort to include citation data ('reference' and 'reference by') and when the amount of data accumulates to certain level, cocitation analysis can be performed and included into our ranking algorithm. Citation can be viewed as important authority or trust vote for certain paper when it gets cited by other papers - 'reference by'. Shared interest can be mined based on the citations - 'reference' of the paper. Both of them can contribute to

No	Paper Title	Author	Score
1	Bringing Semantics to Web Services: The OWL-S Approach.	David L. Martin et al.	229285
2	The Unified Problem-Solving Method Development Language UPML.	Dieter Fensel et al.	198899
3	DAML-S: Semantic Markup for Web Services.	Anupriya Ankolekar et al.	178526
4	A case for automated large-scale semantic annotation.	Stephen Dill et al.	174802
5	Emergent Semantics Systems	Karl Aberer et al.	168973
6	Emergent Semantics Principles and Issues.	Karl Aberer et al.	166527
7	Knowledge Representation on the Web.	Stefan Decker et al.	164742
8	Enabling knowledge representation on the Web by extending RDF schema.	Jeen Broekstra et al.	159794
9	AI for the Web - Ontology-Based Community Web Portals	Steffen Staab et al.	158056
10	Semantic community Web portals.	Steffen Staab et al.	153064
11	OIL in a Nutshell.	Dieter Fensel et al.	151955
12	A new journal for a new era of the World Wide Web.	Stefan Decker et al.	142823
13	IEEE Intelligent Systems	Ian Horrocks et al.	141979
14	Web Services: Been There, Done That?	Steffen Staab et al.	141336
15	On2broker: Semantic-based access to information sources at the WWW.	Dieter Fensel et al.	137839
16	EDUTELLA: a P2P networking infrastructure based on RDF.	Wolfgang Nejdl et al.	134783
17	An Information Food Chain for advanced Applications on the WWW.	Stefan Decker et al.	128618
18	SWAP - Ontology-based Knowledge Management with Peer-to-Peer Technology.	Marc Ehrig et al.	126373
19	Managing RDF Metadata for Community Webs.	Sofia Alexaki et al.	119743
20	Crossing the Structure Chasm.	Alon Y. Halevy et al.	115349

TABLE I

TOP 20 PAPERS IN THE AREA OF 'SEMANTIC WEB' THAT DO NOT CONTAIN THE BIGRAM 'SEMANTIC WEB'

Rank	FoafRank results	Score
1	Dan Brickley	1262558
2	Libby Miller	943774
3	Andreas Harth	360795
4	Martin Dzbor	250619
5	Aaron Swartz	225834
6	Sam Chapman	187833
7	Christian Halaschek-Wiener	171213
8	Stefan Decker	139437
9	Eric Miller	136507
10	Jos de Bruijn	126114
11	Perry Groot	126632
12	Heiner Stuckenschmidt	118125
13	Marta Sabou	113731
14	Frank van Harmelen	109945
15	Mark van Assem	105962
16	Ikki Ohmukai	104048
17	Jeen Broekstra	103167
18	Peter Mika	101896
19	Ronny Siebes	97671
20	Maarten Menken	97618

TABLE II

TOP 20 AUTHORS WITH PUBLICATIONS IN THE AREA OF THE SEMANTIC WEB BASED ON FOAF:KNOWS RELATION

Rank	AuthorRank results	Score
1	Shamkant B. Navathe	22162215
2	Wendy Hall	15338377
3	Farshad Fotouhi	14160971
4	Wesley W. Chu	10659684
5	Elisa Bertino	9638098
6	Chris A. McMahon	7534267
7	Diego R. Lpez	7300322
8	Matthias Jarke	6692724
9	Tharam S. Dillon	6597083
10	Michael Stonebraker	6545999
11	Armin B. Cremers	6241745
12	W. Bruce Croft	6077242
13	John Mylopoulos	5779595
14	Ian T. Foster	5754473
15	Christos Faloutsos	5521122
16	Gio Wiederhold	5381379
17	W. A. Gray	5246268
18	A. Min Tjoa	5024586
19	Michael G. Strintzis	4898049
20	Boi Faltings	4300588

TABLE III

TOP 20 AUTHORS WITH PUBLICATIONS IN THE AREA OF THE SEMANTIC WEB BASED ON DBLP COAUTHOR RELATION

the current ranking algorithm [9]. We want to improve the ranking algorithm to get a similarity measure by clustering the topics (and find the dominant topic inside a cluster to construct subclass relations) based on bigram links for papers. Our future work is to further refine the AuthorRank algorithm and identify the proper weights for the combination of AuthorRank with co-word and co-author analysis [10]. The AuthorRank algorithm can be further broadened to use other metadata, such as RSS and Dublin Core. Some experiments need to set up to test such idea. Visualization can bring end users the friendly interface and better understanding. Especially based on FOAF knows, social network can be portrayed. So discovering useful visualization techniques is also one of our future focuses. Our final goal is to build up a Semantic Web Search Engine (SWSE) based on efficient ranking algorithms.

REFERENCES

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bring order to the web," Standord University, Tech. Rep., 1998.

[2] G. Chowdhury, S. Foo, and W. Qian, "Bibliometric information retrieval system (birs): A web search interface utilizing bibliometric research results," *Journal of the American Society for Information Science*, 2000.

[3] G. Chowdhury and S. Fo, "Mapping intellectual structure of information retrieval: An author cocitation analysis, 1987-1997," *Journal of Information Science*, 1999.

[4] G. Chowdhury and S. Foo, "Bibliometric cartography of information retrieval research by using co-word analysis," *Information Processing and Management*, 2001.

[5] B. Gaume, K. Duvignau, O. Gasquet, and M. Gineste, "Forms of meaning, meaning of forms," *J. Exp. Theor. Artif. Intell.*, 2004.

[6] M. Naaman, H. Garcia-Molina, A. Paepcke, and R. Yeh, "Leveraging context to resolve identity in photo albums," Stanford Database Lab., Tech. Rep., 2005. [Online]. Available: http://dbpubs.stanford.edu:8090/pub/2005-2

[7] S. Decker, "Yars: Optimized index structures for querying rdf from the web," submitted for publication.

[8] G. Grimnes, P. Edwards, and A. Preece, "Learning meta-descriptions of the foaf network," in *Proceedings of the Third International Semantic Web Conference ISWC 2004*.

[9] A. Balmin, V. Hristidis, and Y. Papakonstantinou, "Objectrank: Authority-based keyword search in databases," in *Proceedings of the 30th VLDB Conference*.

[10] B. Aleman-Meza, C. Halaschek-Wiener, I. Budak, C. Ramakrishnan, and A. Sheth, "A flexible approach for analyzing and ranking complex relationships on the semantic web," *IEEE Internet Computing*, 2005.