DISSEMINATING KNOWLEDGE THROUGH TAGS: RECOMMENDING TAGS FOR SCIENTIFIC RESOURCES

Anwar Us Saeed[1, 3], Muhammad Tanvir Afzal[2], Atif Latif[1,4], Klaus Tochtermann[5, 6]

[1]Institute for Knowledge Management(IWM), [5]Know-Center
Graz University of Technology, Inffeldgasse 21a, 8010 Graz, Austria,
{[3]anwar.ussaeed, [4]atif.latif} @student.TUGraz.at,
[6]klaus.tochtermann@TUGraz.at

[2]Institute for Information Systems and Computer Media (IICM)
Graz University of Technology, Inffeldgasse 16c 8010, Graz, Austria
mafzal@iicm.edu

Abstract – **Knowledge diffusion has prime importance for generation of new knowledge. The creation of new knowledge is not possible without referring or consulting the past work. Two very important potential flows related to knowledge diffusion can be observed in the common practice of researchers. First, in scientific research knowledge diffusion estimation using citation counts is generally used to establish the value of knowledge which inflates citations. Second the researchers use cited work to search the connected and related resources. Recently the social and collaborative phenomena termed as Web 2.0 has spurred new era of knowledge and information flow on the web. Its potential for the growth and diffusion of scientific knowledge has not been well explored. The emerging social and collaborative applications, such as tagging and bookmarking, are transforming the ways scientists and researchers organize their personal and collaborative information spaces. These bookmarking and tagging applications provide open data and rich metadata resources such as tags. Past research shows that the bookmarking and tagging can be used as a supplementary indicator for measuring research popularity and knowledge diffusion. However the current work exploits author keywords of scientific publications to link these resources with relevant tags extracted from a social bookmarking application such as CiteULike. This work compares, for a focus resource, the tags extracted from CiteULike based on author keywords with their corresponding tag cloud of CiteULike. The result shows that system extends the authors keyword set with social tags providing links to rich and focused resources in CiteULike. This also enhances the serendipitous discovery of emerging concepts related to the focused resource. Such a system may enhance the discovery of related and popular resources for researchers. This dataset has been made available publicly for scientific community.**

Keywords: **Knowledge diffusion, social bookmarking, tagging, recommendation system.**

1. INTRODUCTION

Knowledge is of prime importance for economic and social development. The diffusion of knowledge holds an important role in the creation and distribution of knowledge boons. The diffusion of published (codified) scientific knowledge has been mainly investigated in the past to study the structures and properties of knowledge diffusion in scientific domain. In science and technology citations are considered as an indicator for volume of diffusion of a published work. Citation is a relationship between two published papers or articles where normally the author(s) of 'citing' paper infer(s) from and refer(s) to the part of 'cited' paper used to extend or create new knowledge published in the 'citing' paper. Citations are also used to measure the impact of research. It is considered that, to some extent, collaborative behavior may affect the citations of a paper or an article. Usually researchers collaborate and jointly report in their research publications the new ideas and findings of research are established after conversations among them. When more than one authors share a published work, they are called coauthors. Co-authorship analysis and citation analysis are the popular techniques used to assess diverse aspects of knowledge, in science and technology. Knowledge diffusion in general is analyzed using diffusion of innovations, epidemiology, collaboration network analysis (co-authorship analysis) and citation analysis techniques.

In addition to the study of the diffusion of (codified) scientific knowledge through citations, the need of web based indicators for assessment of different aspects of science and technology has also been pointed out in (Scharnhorst and Wouters, 2006) (Day, 2008). The latest developments in the Web termed 'Web 2.0' or 'Social Web' has provided access to open source data and metadata resources. Kleinberg argues that the web will 'bring evolution in future in the ways of scientists' work and their communication' (Kleinberg, 2004). Furthermore, the recent trends of contributory web and inflated web-based publishing have the potential to blur the boundaries of formal and informal scientific communications. The applications like the 'Encyclopedia of Life' (EOL) may become very popular future publishing platforms for scientists (Us Saeed et al., 2007). Every day, the research work is getting more and more convoluted with the emerging structures of web. It is feared that the dynamics of diffusion of scientific literature on the web in future may not be assessable by conventional techniques alone. This emphasizes the need for a particular type of web indicators, one of which may be bookmarking /tagging, which are within the streams of this new form of web evolution. This research intends to explore the potentials of these bookmarking applications in the diffusion of knowledge and its estimation. Tagging practices have an added advantage to augment the understanding of knowledge diffusion by providing an additional element – the user context in tagging a resource of knowledge (to understand the better reason about the usage of knowledge).

Past research (Us Saeed et al., 2008a) shows that bookmark counts in CiteULike mines the interest of researchers in a particular scientific resource. The bookmark counts are correlated positively with the citations of that resource. This result can be used to establish the popularity and hence citation count or quality of that resource. The research also concludes that the tag terms assigned by users to a particular scientific paper of WWW'06, in social bookmarking applications, frequently re-occur in the titles of its citing papers. This shows that tag terms hold the diverse context of diffusion of a scientific research.

Citation count also inflates diffusion by increasing popularity of research and is also considered as an indicator for establishing the quality of research. Along with this, the researchers also use citations or references to search the connected and related resources hence increasing diffusion of interlinked knowledge. Based on these potential uses of citations in the research community and results of past research, which shows that citation count and bookmark counts are positively correlated, we argue here that bookmark counts of research papers can be used in a similar way as an alternative popularity indicator. Along with this we propose a tag recommender system for scientific papers. These recommended tags provide a link to the most related resources which gives two benefits. 1) these resources will be directly related to the content and context of diffusion of that paper which is implicitly derived from the tags extraction mechanism 2) the researcher can explore the interlinked and related resources as they use references or citations.

The tag recommender system exploits author keywords of scientific publications to link these resources with tags in CiteULike which is a social bookmarking and tagging application. We also compared, for a focused resource, the tags extracted from CiteULike based on author keywords with their corresponding tag cloud of CiteULike. The result shows that system extends the authors keyword set with social tags providing links to rich and focused resources in CiteULike. This also enhances the serendipitous discovery of emerging concepts related to that resource. Such a system may enhance the discovery of related and popular resources for researchers. This paper contributes to the knowledge diffusion discussion by exploring the potential of tagging and bookmarking in scientific knowledge diffusion.

2. SOCIAL BOOKMARKING SYSTEMS AND THEIR POTENTIAL IN MEASURING KNOWLEDGE DIFFUSION

Social bookmarking and tagging has become a very successful phenomenon in the web and getting more popular day by day. Systems adhering to these principles transformed the way about managing and dissemination of content in the conventional web environment. These systems enable the users to add keywords (tags) to web resources (web-pages, images, documents, papers) without having to rely on a controlled vocabulary (Marlow et al., 2006). It's potential to improve the search on the web, resulted in new forms of social communication and generated new opportunities for data mining. However, in our previous study we found that tagging system got real recognition and rated as an integral part of Web 2.0 after year 2005. We investigated bookmarking and tagging as a medium to measure the knowledge diffusion. However, there also lies some reservations of research community in considering these systems as a supplementary measure for knowledge diffusion. One of them is their inability to have control on the users for specifying relevant tags to the resource and handling manipulation of these tags to various contexts. This claim can be true for tagging non scientific content but our previous experimental findings revealed that most users do tag a document only after having some understanding of the content and use them in their particular personal context. Meanwhile, for sure some further efforts may be needed to enhance the tagging applications to make them more strict systems for managing tags in a controlled way. One approach adopted here in the proposed recommendation system for filtering the tags with the author key words as seeds can also be effective to resolve this problem.

In the fields of emergent semantic (Mika, 2005), Information Retrieval (Wu et al., 2006) (Hotho et al., 2006) and user profiling (Huang et al., 2008) tagging is considered as a driving component. (Michlmayr et al., 2007).Wu and colleagues (2006) have shown that "In a collaborative tagging system, tags codify the knowledge of relationships among documents and concepts represented by the tags. Harvesting individual through folksonomies therefore can benefit the whole society." (Wu et al., 2006). Mika (2005) has studied the tagging behaviors and their usage in del.icio.us, as an emerging bookmaking service. He used actor, concept, and instance nodes as a tripartite graph to explain the emergence of ontologies from social context where he considers tags as a socially represented concept.

In this study, we intend to compare the tagging behaviors with the knowledge diffusion mechanisms and their corresponding contexts. We also use them for effective tags and resource recommendation for scientific papers. Literature has shown that 'context' became an important consideration in any discussion of codified knowledge (Cowan et al., 2000). However, in previous works there were very limited explicating instances about the usage of context in diffusion studies. For example, Tsai (2001) described the contextual flow of knowledge within scope of an organization, and Chen and colleagues (2007) used context in the geospatial distribution of diffusion.Heterogeneity of context in reuse of knowledge implies the need for an indicator in which the constituent parts can be rendered commensurably. Tags may augment the context of the knowledge being used by different users (Wu et al., 2006). We have shown in Figure 2 that how tagging can be used to contextualize the knowledge diffusion.

Previously many constructs has been employed to measure the Knowledge diffusion, one of the popular and important one is Citations. Citations are studied in different ways like scientific fronts, a service provided by ISI since Feb 2008 which performs a co-citation analysis within different subfields of a broad subject. They built subfields by extracting keywords from titles of highly co-cited papers. But there is a lack of a standard taxonomy for a particular field. For example if we want to study subfields for computer science, one may suggest that ACM standard taxonomy can be used, but research has shown that a large amount of documents in digital libraries are not categorized according to this taxonomy and then mapping of papers to this classification becomes problematic when the paper is not explicitly stated into a particular category which is the case in most of the papers (Cameron et al., 2007).Previous research showed that there are certain limitations of citations like 1). citations of existing papers do not necessarily mean that the cited-by paper is regenerating knowledge by using knowledge from the cited papers 2) Citations inability to highlight the real context of the citing paper for example citations are made to just give a broad level background study and the context of cited paper is not always clear by reading the citing paper. 3) Citation analysis may not always predict the contextual use of the knowledge 4) Limitation of citations to just understand the codified knowledge. For example in the case of applied research, knowledge is not often used to create new knowledge, thus receives a fewer citations but is used practically in various fields. This knowledge for practice, however, cannot be measured by citations.

By taking these limitations in account, we have proposed that bookmarking/tagging got a potential to be used as a supplementary measure in predicting and estimating the contextualized knowledge diffusion. We think that tagging may tackle the situation in a more convincing way as compared to citations because tags are explicitly specified by the users in their own context when viewing a particular paper. For example a user tags a particular paper most of the time as "Web 2.0", but at the same time other contexts of users for that particular paper will also be a part of its tag cloud. As investigated by Mika (2005), these tags and their proportional percentages can be used to make an automatic taxonomy.

We explore the potential of bookmarking and tagging with our safe assumption, that people tag something: 1) if they conceptually understand the content and 2) if they perceive it to be useful in their own context (of work).

3. EMPIRICAL STUDY OF RELATIONSHIP OF BOOKMARK COUNTS AND TAG TERMS WITH CITATIONS

We performed an exploratory case study (Us Saeed et al., 2008a; Us Saeed et al., 2008b). We analyzed the published 84 papers of the conference World Wide Web 2006 (WWW'06). The WWW'06 was selected as a dataset because of its special focus and popularity. The papers presented at the WWW conference series generally discuss the future evolution of the web. That is why we were expecting to find WWW papers both frequently cited and tagged in social bookmarking applications. The higher numbers of citations show the large scale of volumetric knowledge diffusion and high impact of scientific resources. The citation ranks for research papers are usually predicted using various factors. These factors include multi-author publications, geographical positions of co-authors, co-authors' network, and multi-institutional involvement in a publication. However, with the evolution of the Web 2.0, bookmarking and tagging applications are considered as the popularity measure for scientific resources. As our focus of study was to compare different citation prediction models, we need a dataset of research papers from a conference which is popular and within a particular focus related to the web (so that the potential research community is already integrated within the bookmarking systems). Considering all of these factors, we selected one of the most highly ranked conference i.e. World Wide Web conference.

We took the event from the year 2006, because tagging applications were not popular before the year 2006. The assumption was that a certain degree of popularity would be required for representing real tagging behaviors. We did not select the event from 2007 or 2008 because normally it takes 1-2 years to enable the regeneration of the new knowledge.

We explored the selected papers in three common social bookmarking and tagging systems CiteULike[8], BibSonomy[9] and del.icio.us[10]. Although BibSonomy and del.icio.us give access to their search APIs, yet our initial experiments showed that searching a particular paper which have some special characters (like : , - _ ‘ “ & vs. / etc.) in its title does not find its match in the tagging application. It was found that sometime the same user (who tags a resource) is listed repetitively for one paper in these applications. It was also found that sometimes same user tags the same paper with different tags in different times. This leads to miscount of the total number of users for a paper. By considering all of these limitations, we safely explored the bookmark counts, tags and the users in these applications. Citations were acquired from Google scholar[11] manually because Google Scholar does not provide open access API to explore the citations. We tabulated the dataset year wise from bookmarks/tags and citations with the paper numbers as ‘ids’ and their titles extracted from WWW‘06 website[12]. The ids are maintained in the order of paper titles listed on the website. Figure 1 depicts various modules of the study design for the research.

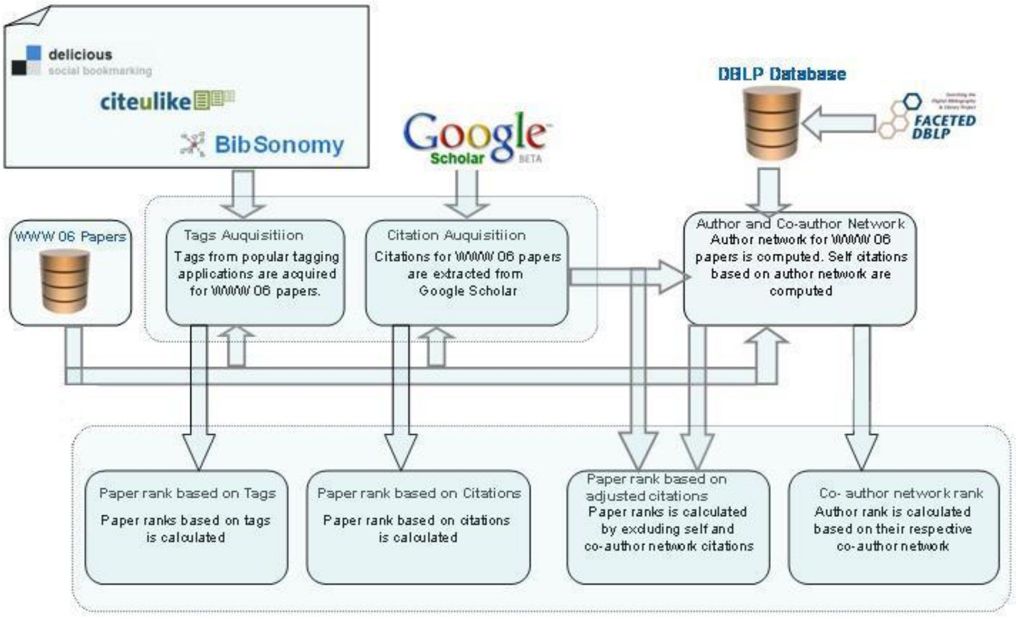

delicious
social bookmarking
citeulike
BibSonomy
Google
Scholar
BETA
DBLP Database
FACETED DBLP
WWW 06 Papers
Tags Auquisitiion
Tags from popular tagging applications are acquired for WWW 06 papers.
Citation Auquisitiion
Citations for WWW 06 papers are extracted from Google Scholar
Author and Co-author Network
Author network for WWW 06 papers is computed. Self citations based on author network are computed
Paper rank based on Tags
Paper ranks based on tags is calculated
Paper rank based on Citations
Paper rank based on citations is calculated
Paper rank based on adjusted citations
Paper ranks is calculated by excluding self and co-author network citations
Co- author network rank
Author rank is calculated based on their respective co-author network

Figure 1. System design

Below we explain how were the data sets for bookmarks, citations, co-authors’ network acquired prior to computing different citation prediction models.

Tags and bookmarks for WWW‘06 papers were collected from the CiteULike, BibSonomy and De1.icio.us based on their popularity in the Web research community. The total bookmarks for the 84 papers were 1051. Citations for WWW‘06 papers were acquired using Google Scholar. Although Google Scholar does not provide a search API for citation extraction, but Google Scholar was chosen because of its large index. Google Scholar index covers "peer-reviewed papers, theses, books, abstracts and articles, from academic publishers, professional societies, preprint repositories, universities and other scholarly organizations" (About Google Scholar, 2009). Google Scholar also finds some false positive citations like citations to press releases, resumes, and links to bibliographic records for cookbooks (Price, 2004). But we have safely extracted all citations manually for WWW‘06 papers. The total citations for the 84 papers were 1165.

[8] http://www.citeulike.org/

[9] http://www.bibsonomy.org

[10] http://del.icio.us/

[11] http://scholar.google.com/

[12] http://www2006.org/

3.1 Author's and co-authors' network

The citation rank studies are usually based on co-authors' network. We computed citation rank for WWW'06 papers based on a number of bookmarks and co-authors' network. To build a co-authors' network, we selected a dataset of DBLP++ (Diederich et al., 2007). This is an enhanced dataset of DBLP (a digital library for computer science publications). DBLP indexes WWW'06 conference in particular and contains 1,048,576 publication records in general. DBLP is managed manually. Due to this, it does not include the inherited problems of autonomous systems. This module performs four tasks:

1) Finds authors of papers of WWW'06 conference. 2) Finds citing authors for all papers of WWW'06. 3) Computes a co-authors' network based on the original authors of the paper. The co-authors' network is computed up to 2 degrees of separation. The average co-authors' network for WWW'06 authors was 119. 4) Computes self citations and citations by a co-author's network.

As already mentioned that there were 1165 overall citation for WWW'06 conference papers. Self citations were 208, citations in the first level co-authors' network were 60 and citations in the second level co-authors' network were 26. These figures also indicate that self citations and citations in co-authors' network (up to 2 levels) accumulatively were only 25% of all citations.

3.2 Findings from the Study

3.2.1 Bookmark counts positively correlates to citations

In the initial state of our study, we found a positive correlation ($r = 0{,}65$, $p = 2.133e\text{-}11$) between the total number of bookmarks and the total number of citations from May 2006 to May 2008 for all the papers. This finding indicates that the bookmarking and tagging behavior somehow matches with the citation behavior.

3.2.2 Bookmarking may have the potential to foretell the future volume of knowledge diffusion

We calculated the average number of users in table 1 by adding all the users from three tagging applications for a particular paper and dividing it by three (i.e. number of tagging applications). We observed that if the average is higher than 6, then the tagged paper also gets reasonable number of citations (≥ 7). See table 1. For such papers the major number of citations came from the year 2007. However, for the same papers, the major number of user's bookmark counts came from the year 2006.

This is logical, because the bookmarks/tags will come earlier in time than the citations. The regeneration of knowledge needs more time than the selection of a piece of knowledge. This makes the case interesting for tagging analysis, because it shows a possible potential of the bookmark counts to forecast the future volume of knowledge diffusion.

Table 1. Heavily bookmarked papers in 2006 got heavy citations in 2007

Paper ids	Avg. No. of users per tagging application (> 6)	Total user bookmark counts (06)	Citations in 2006	Citations in 2007	Total citations
9.	7	7	11	44	61
10.	8	20	3	6	12
17.	9	13	4	11	18
23.	49	80	9	37	49
24.	11	18	5	15	23
25.	7	14	1	19	23
31.	7	7	1	7	8
50.	40	100	10	24	43
51.	32	37	4	32	39
69.	30	41	34	68	112
73.	21	21	5	24	33

3.2.3 Tagging may have the potential to foretell the context of future knowledge diffusion

A lightweight tool was developed to create tag-clouds. Using this tool, we created two tag-clouds for each paper: 1) Tag-cloud of the tag terms from all tagging applications. 2) A second tag-cloud was generated by selecting the matched tag terms of first tag-cloud in the titles of the respective citing paper. The font size of second tag-cloud is assigned on the matching frequency of the terms in the titles of citing papers. The trend for heavily tagged and cited papers is visualized in Figure 2.

The results showed that about 16 to more than 22 percent tagged terms matched with the title terms of the citing papers. This result is in line with our assumption that tagging may forecast the context of knowledge diffusion. We found that the bigger portion of the tags represent the content of the paper being tagged, while the rest represents the context of future use.

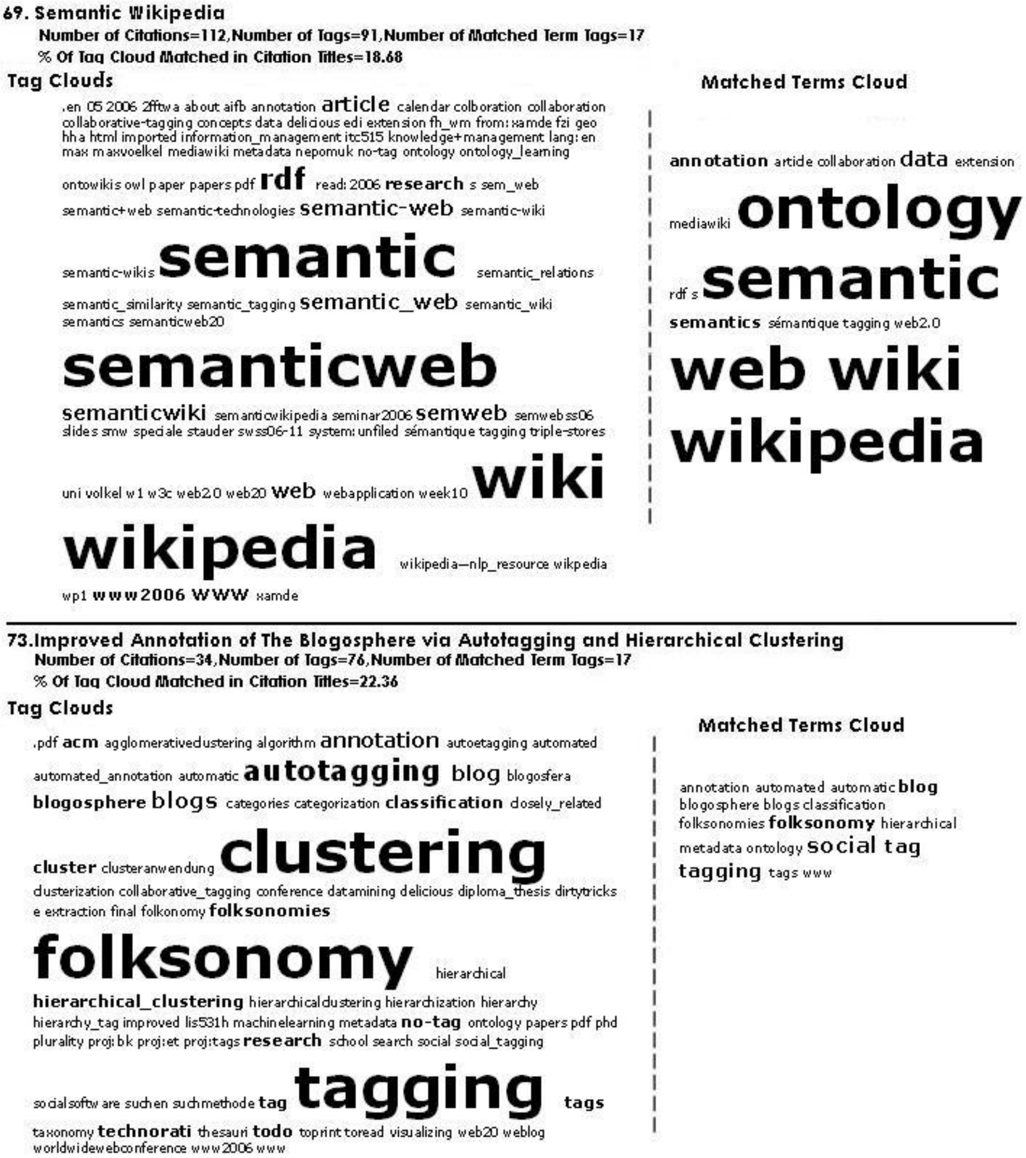

69. Semantic Wikipedia
Number of Citations=112,Number of Tags=91,Number of Matched Term Tags=17
% Of Taq Cloud Matched in Citation Titles=18.68
Tag Clouds
.en 05 2006 2fftwa about aifb annotation article calendar colboration collaboration collaborative-tagging concepts data delicious edi extension fh_wm from:xamde fzi geo hha html imported information_management itc515 knowledge+management lang:en max maxvoelkel mediawiki metadata nepomuk no-tag ontology ontology_learning ontowikis owl paper papers pdf rdf read:2006 research s sem_web semantic+web semantic-technologies semantic-web semantic-wiki semantic-wikis semantic semantic_relations semantic_similarity semantic_tagging semantic_web semantic_wiki semantics semanticweb20 semanticweb semanticwiki semanticwikipedia seminar2006 semweb semwebss06 slides smw speciale stauder swss06-11 system:unfiled sémantique tagging triple-stores uni volkel w1 w3c web2.0 web20 web webapplication week10 wiki wikipedia wikipedia—nlp_resource wikpedia wp1 www2006 WWW xamde
Matched Terms Cloud
annotation article collaboration data extension mediawiki ontology rdf s semantic semantics sémantique tagging web2.0 web wiki wikipedia
73.Improved Annotation of The Blogosphere via Autotagging and Hierarchical Clustering
Number of Citations=34,Number of Tags=76,Number of Matched Term Tags=17
% Of Taq Cloud Matched in Citation Titles=22.36
Tag Clouds
.pdf acm agglomerativeclustering algorithm annotation autoetagging automated automated_annotation automatic autotagging blog blogosfera blogosphere blogs categories categorization classification closely_related cluster clusteranwendung clustering clusterization collaborative_tagging conference datamining delicious diploma_thesis dirtytricks e extraction final folkonomy folksonomies folksonomy hierarchical hierarchical_clustering hierarchicalclustering hierarchization hierarchy hierarchy_tag improved lis531h machinelearning metadata no-tag ontology papers pdf phd plurality proj:bk proj:et proj:tags research school search social social_tagging socialsoftware suchen suchmethode tag tagging tags taxonomy technorati thesauri todo toprint toread visualizing web20 weblog worldwidewebconference www2006 www
Matched Terms Cloud
annotation automated automatic blog blogosphere blogs classification folksonomies folksonomy hierarchical metadata ontology social tag tagging tags www

Figure 2. Tag cloud comparison of heavily cited and tagged papers.

3.2.4 Paper rank models

Bookmarks, citations and co-authors' network are further used to establish different models for paper rank.

a) Paper rank based on bookmarks

This model ranks papers based on their popularity on Web (tagging and bookmarking applications), the number of users who bookmarked a paper are aggregated from different applications to form a total user count for a particular paper. The large number of users ranks a paper on top in this model.

b) Paper rank based on citations

This model ranks papers based on their citation counts. The extracted citations are used to rank paper in this model. The high number of citations ranks a paper on the top in this model.

c) Paper rank based on adjusted citations

As mentioned earlier there are some previous studies which talk about the adjustment of scientific impact based on co-authorship and its network. There is a need to adjust the citations by excluding self citations and citation loops (Ioannidis et al., 2008). There is evidence that, to some extent, sharing of self citations may be inflated by co-authorship (Glänzel and Thijs, 2004).

d) Co-authors' network rank

In this model, we computed the co-author network for all authors of WWW'06 conference. Author's network is computed up to 2 levels. An author is selected for each publication in WWW'06, his co-authors' count is added form the author's network count. Furthermore 2nd level of coauthors' count is also added to the original author's network count. In this way, the author's network count is calculated for each author of WWW'06 conference. Authors are ranked based on their respective co-authors' count. All authors' network counts for a particular publication are added to form the absolute count for a paper. This model assumes that the papers with high number of authors' and coauthors' count will receive high citations and hence the higher rank.

3.2.5 Citation Ranks Prediction Models

Based on the collected bookmarks, citations and co-authors' network for WWW'06 conference papers, we have explored citation rank model by applying different variables and then compared the results. We have applied linear regression analysis. Linear regression is a form of regression analysis in which the relationship between one or more independent variables and another variable, called dependent variable, is modeled by a least squares function, and represented by a Linear Regression (LR) equation. The details of citation rank model based on different variables are depicted below.

a) Citation rank prediction model based on bookmarks

In this model bookmarks are used as an independent variable while citations are taken as a dependent variable. The linear regression equation model is as follows:

$$0.69 * \text{variable (bookmark - rank)} + 6.21 \qquad (1)$$

In the model equation (1) 0.69 is called the regression coefficient. It explains the behavior of change in the value of dependent variable for small change in bookmark rank. The term 6.21 is called the disturbance or noise term.

b) Citation rank prediction model based on co-author network

In this model co-author's network (calculated in section 3.2.4) is used as an independent variable while citations are taken as a dependent variable. The linear regression equation model is as follows:

$$0.46 * \text{variable (coauthor rank)} + 30.27 \qquad (2)$$

In the model equation (21) factor 0.46 is called the regression coefficient. It explains the behavior of change in the value of dependent variable for small change in co-author counts. The term 30.27 is called the disturbance or noise term.

c) Citation rank prediction model based on adjusted citations

In this model bookmarks are used as an independent variable while citations are taken as a dependent variable. The citation counts are adjusted by excluding self citations. The linear regression equation model is as follows:

$$0.69 * \text{variable (bookmark rank)} + 6.85 \qquad (3)$$

In the model equation (3) factor 0.69 is called the regression coefficient. It explains the behavior of change in the value of adjusted citation rank for small change in bookmark rank. The term 6.85 is called the disturbance or noise term.

The correlation coefficient established on WWW‘06 papers by bookmarking count model is 0.6003 which is considered as a fair correlation, while it is 0.1559 by co-authors’ network model. This is not so good. This correlation coefficient is enhanced up to 0.6657 by excluding the self citations

The mean absolute error is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. It was 5.3727 by bookmark model while this mean error was much higher (18.1428) in co-authors’ network. This error is reduced up to 4.3821 with the self citation adjustment.

Our results have proved that citation rank prediction based on bookmark ranks of papers have got fairly good results than co-author network model (see Table 3). The citation loops like self citations are considered in this research (see Table 2). This furthermore improves the correlation coefficient and reduces the mean absolute error (see Table 4). However, these results are obtained for WWW‘06 conference papers and further studies are necessary to their generalization.

Table 2. Top 5 Ranks of Papers with respect to bookmarking and their respective other Ranks

Paper ID	Bookmark Rank	Citation Rank	Adjusted Citation Rank
23	1	3	3
50	2	5	7
51	3	6	5
69	4	1	1
73	5	7	6

Table 3. Top 5 Ranks of Papers with respect to bookmarking and their respective citation Ranks

Paper ID	Paper Rank based on coauthor count	Citation Rank
49	1	6
23	2	3
50	3	5
69	4	1
65	5	26

4. RECOMMENDING TAGS FROM CITEULIKE

In previous sections, it has been shown that there exist a positive correlation between bookmark counts and citations. A paper starts getting tags from the users of the social bookmarking system immediately after its publication. This section explains how scientific papers can get relevant resources (tags and papers) for papers published within digital journals or liberaries. For this exercise, we have focused on WWW‘06 as a source data set. The social bookmarking system used in our experiments was CiteULike. The CiteULike is a social bookmarking system where a huge number of users share scientific papers and tag them accordingly. Our task is to find the most relevant resources from CiteULike for all papers published within WWW‘06. On the WWW‘06 side, every paper is assigned with suitable keywords by the authors of the paper, while on CiteULike side, papers are tagged with some keywords by the users of the CiteULike. To find relevant resources for WWW‘06 papers from CiteULike, we used authors’ assigned keywords and compared them with CiteULike tags. The papers at WWW‘06 are further annotated with the matched tags. Furthermore, the tags are pushed to users by looking to their local context and tasks at hand.

Table 4. Comparison of citation prediction models based on LR

LR	Prediction model based on bookmark rank	Prediction model based on Co-author network	Prediction model based on adjusted citations
Correlation coefficient	0.6003	0.1559	0.6657
Mean absolute error	5.3727	18.1428	4.3821
Root mean squared error	6.6213	20.8102	5.5976
Relative absolute error	75.6676 %	99.4605 %	71.1488 %
Root relative squared error	79.9746 %	98.7775 %	74.6248 %
Total Number of Instances	84	84	84

4.1 WWW‘06 dataset

This dataset is comprised of all published papers in the conference World Wide Web 2006.

Total papers published in WWW‘06 = 84
Total Keywords for all papers = 5129
Unique Keywords = 107

4.2 CiteULike dataset

The dataset of CiteULike we used was acquired in August, 2009. The statistics for tags and papers is shown below.

Total tag assignments in CiteULike = 6.5 million
Total Papers in CiteULike = about 2 million
Unique tags = 348420

4.3 Matching author’s keywords with CiteULike tags

To match papers’ keywords of WWW‘06 with CiteULike tags, a two-tier approach was adopted. First we tried to find an exact match between papers’ keywords and CiteULike tags. Subsequently, a partial match between both datasets was checked. The partial match enhanced discovery of relevant tags but also introduced some noise. Afterwards, some heuristics were used to clean the noise and the discovered tags were used to annotate the corresponding papers.

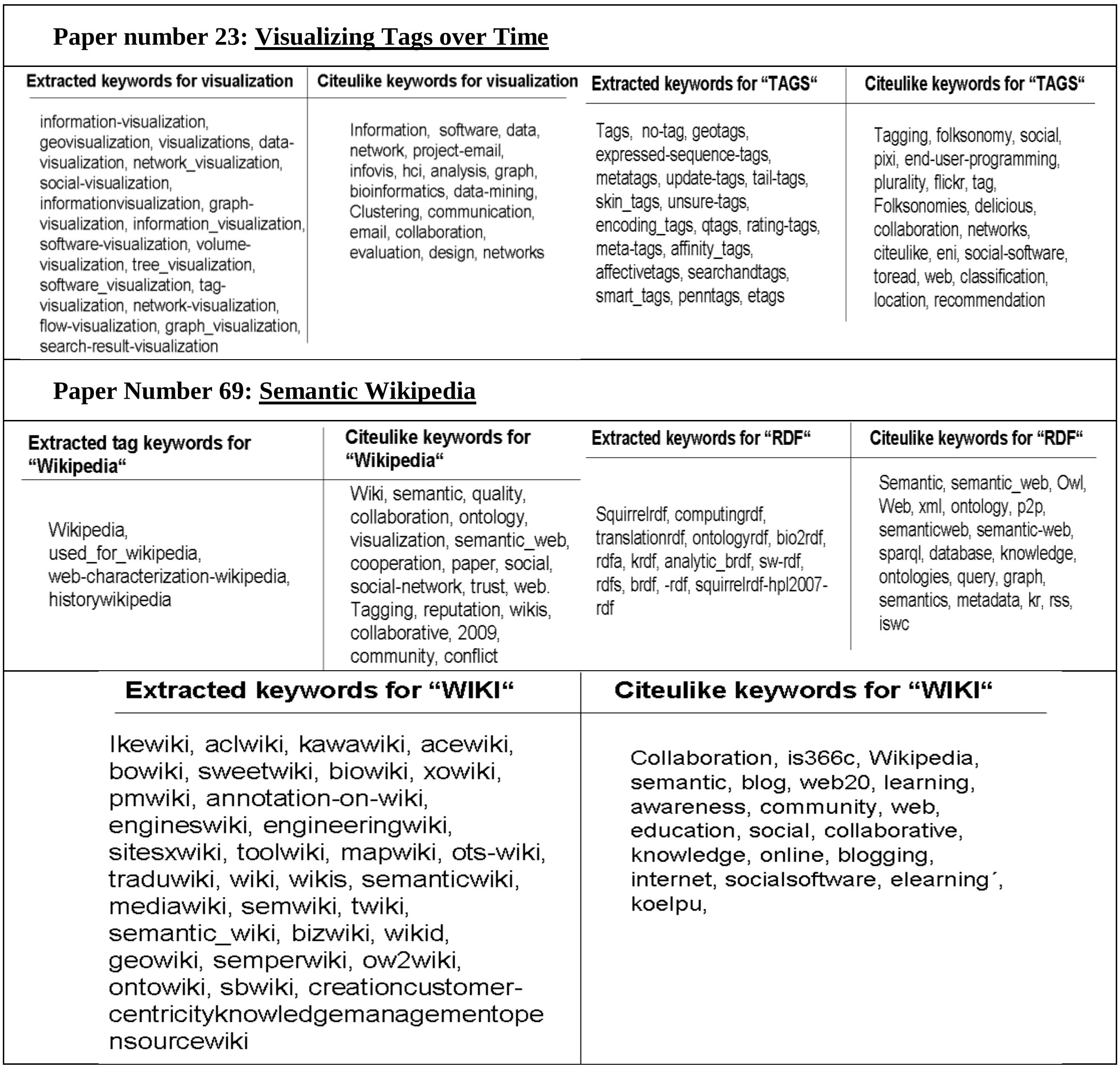
1) Direct Match

WWW papers for which at least one keyword is matched= 52/84 = 62%
Unique Keywords of WWW‘06 matched = 102/107 = 95%

2) Partial Match

WWW‘06 Papers for which at least one tag is matched = 52/84 = 62%.
Total results of WWW‘06 Keywords matched with CiteULike = 5129
Total CiteULike unique tags matched = 4228/348420

In the direct match, the system found one exact tag from CiteULike for each of 102 unique keywords of WWW‘06. The knowledge discoveries are significantly enhanced by employing partial match. The partial match found a total of 5129 matching tags from CiteUlike. This becomes a basis for recommending relevant tags for the focused paper. The partial match enhances the system discoveries significantly for example, the author keyword ‘visualization’ has found its match in the related popular concepts (GeoVisualization, DataVisualization, NetworkVisualization, SoftwareVisualizatuion, GraphVisualization, TreeVisualization, etc).



Paper number 23: Visualizing Tags over Time

Extracted keywords for visualization	Citeulike keywords for visualization	Extracted keywords for “TAGS“	Citeulike keywords for “TAGS“
information-visualization, geovisualization, visualizations, data-visualization, network_visualization, social-visualization, informationvisualization, graph-visualization, information_visualization, software-visualization, volume-visualization, tree_visualization, software_visualization, tag-visualization, network-visualization, flow-visualization, graph_visualization, search-result-visualization	Information, software, data, network, project-email, infovis, hci, analysis, graph, bioinformatics, data-mining, Clustering, communication, email, collaboration, evaluation, design, networks	Tags, no-tag, geotags, expressed-sequence-tags, metatags, update-tags, tail-tags, skin_tags, unsure-tags, encoding_tags, qtags, rating-tags, meta-tags, affinity_tags, affectivetags, searchandtags, smart_tags, penntags, etags	Tagging, folksonomy, social, pixi, end-user-programming, plurality, flickr, tag, Folksonomies, delicious, collaboration, networks, citeulike, eni, social-software, toread, web, classification, location, recommendation

Paper Number 69: Semantic Wikipedia

Extracted tag keywords for “Wikipedia“	Citeulike keywords for “Wikipedia“	Extracted keywords for “RDF“	Citeulike keywords for “RDF“
Wikipedia, used_for_wikipedia, web-characterization-wikipedia, historywikipedia	Wiki, semantic, quality, collaboration, ontology, visualization, semantic_web, cooperation, paper, social, social-network, trust, web. Tagging, reputation, wikis, collaborative, 2009, community, conflict	Squirrelrdf, computingrdf, translationrdf, ontologyrdf, bio2rdf, rdfa, krdf, analytic_brdf, sw-rdf, rdfs, brdf, -rdf, squirrelrdf-hpl2007-rdf	Semantic, semantic_web, Owl, Web, xml, ontology, p2p, semanticweb, semantic-web, sparql, database, knowledge, ontologies, query, graph, semantics, metadata, kr, rss, iswc

Extracted keywords for “WIKI“	Citeulike keywords for “WIKI“
Ikewiki, aclwiki, kawawiki, acewiki, bowiki, sweetwiki, biowiki, xowiki, pmwiki, annotation-on-wiki, engineswiki, engineeringwiki, sitesxwiki, toolwiki, mapwiki, ots-wiki, traduwiki, wiki, wikis, semanticwiki, mediawiki, semwiki, twiki, semantic_wiki, bizwiki, wikid, geowiki, semperwiki, ow2wiki, ontowiki, sbwiki, creationcustomer-centricityknowledgemanagementopensourcewiki	Collaboration, is366c, Wikipedia, semantic, blog, web20, learning, awareness, community, web, education, social, collaborative, knowledge, online, blogging, internet, socialsoftware, elearning´, koelpu,

Figure 3. Comparison of recommended tags for particular author keywords and their relevant CiteULike tags

5. RECOMMENDING RELEVANT TAGS FOR RESEARCH PAPERS

The contribution of this research can be structured into two aspects: 1) Discovery of focused set of tagged resources in social bookmarking applications. 2) leads to serendipitous discoveries of relevant and evolving concepts.

The intention of this research is to discover and recommend a set of most relevant and focused tags from social bookmarking applications for scientific resources. It is a common practice of researchers to explore the resources through interlinked chains as through references or citations. The socially annotated libraries like CiteULike also provide an interlinking of resources by using hyperlinked tags. For example CiteULike provides a list of tag terms for a user search keyword ‘visualization’ as shown in Figure 3 (only top 20 are shown). These terms are computed from the tag co-occurrence. e-g. terms related to ‘visualization’ search keyword are the terms which same users assigned to resources along with tag term ‘visualization’. This tag list is organized on the basis of frequency of term occurrence in CiteULike. From the Figure 3, if a user want to explore further resources from CiteULike related tag terms of visualization search keyword, say by clicking on Clustering tag in the list, then the user will get a list of all resources annotated with tag term ‘Clustering’. There might be some resources related to main focus (visualization) somewhere in the list but the returned recourses will be sorted based on clustering keyword rather than visualization keyword which put an extra burden on user to find focused resources. However in our case, we extracted the tag terms from CiteULike tags based on direct and partial match of authors' keywords of a particular research paper. In this way the highly relevant discovered tags are linked with the paper. For example for the WWW'06 paper ' Visualizing tags over time' authors provided keywords are 'visualization', 'tags', 'flickr', temporal evolution' and 'interval covering'. We compared these in CiteULike tags by using direct and partial match. The extracted tag terms for 'visualization' and 'tags' are shown in Figure 3. The extracted tags for visualization remains in the same focus and will link the resources in CiteULike which will often be related to the scope of visualization. Now if a user visit this paper he will see these related tags organized according to author keywords as hyperlinks. For further navigation if a user selects any tag from the extracted list, he/she is directed to the associated resources in CiteULike.

The second contribution of this research is an overall extension of the author keyword concepts into their different subfields and application areas along with some serendipitous discoveries of relevant or evolving concepts. It is obvious from the Figure 3 that the tags extracted for keyword 'visualization' are its subfields like data-visualization, its application areas like network-visualization and evolving concepts like social visualization. This list of keywords signifies an overall picture of popular research in related fields within the focus of a research paper.

6. CONCLUSION AND FUTURE WORK

In this research we discovered a relationship between bookmarks/tags and citations. The case study shows that there exist a positive correlation between bookmark counts and citations. Tag terms also reoccur in the titles of the citing papers. Furthermore, the ranking of papers based on bookmark counts can predict citation counts better than the co-author network.

Afterwards, we found that there are some tags which only show the context of future diffusion but a high percentage of tags show the content of the paper. We linked WWW'06 papers with CiteULike papers. For this purpose, we used authors' assigned keywords to WWW'06 papers and found relevant tags from CiteULike by direct and partial match. The system was able to recommend popular tags for WWW'06 papers and a user had an option to find other relevant resources (papers) that are annotated with the same or similar tag. The result shows that system extends the authors keyword set with social tags providing links to rich and focused resources in CiteULike. This also enhances the serendipitous discovery of emerging concepts related to the focused resources. The dataset for tags recommendation has been made available at http://www.student.tugraz.at/anwar.ussaeed/datasets.html.

REFERENCES

Anjewierden, A., de Hoog, R., Brussee, R. & Efimova, L. "Detecting knowledge flows in weblogs", in 13th International Conference on Conceptual Structures, University of Kassel, Kassel, 1-12, July 2005.

Bettencourt, L M. A., Castillo-Ch´avez,, C., Kaiser, D., Wojick, D.E. Report for the Office of Scientific and Technical Information:Population Modeling of the Emergence and Development of Scientific Fields; http://www.osti.gov/innovation/research/diffusion/epicasediscussion_lb2.pdf ; October 4, 2006

Branstetter, L., "Measuring the impact of academic science on industrial innovation: the case of California's Research Universities". Columbia Business School Working Paper, 2003.

Chen, C., Maceachren, A., Tomaszewski, B., MacEachren, A., "Tracing conceptual and geospatial diffusion of knowledge", in LNCS 4564, pp.265-274, 2007.

Cowan, R., Paul A. David, and Dominique Foray "The Explicit Economics of Knowledge Codification and Tacitness", in Industrial and Corporate Change, 9(2), pp.211-253, 2000.

Day, M.: "Institutional Repositories and Research Assessment", Supporting Study No. 4, UKOLN, ePrints UK Project, Bath, available at: www.rdn.ac.uk/projects/eprints-uk/docs/ studies/rae/rae-study.pdf (accessed 24 April 2008).

Garfield, E. "The epidemiology of knowledge and the spread of scientific information.", Current Contents 35, pp. 5-10 , 1980

Glänzel, W., Thijs, B., "Does co-authorship inflate the share of self-citations?", Scientometrics, Volume 61, Number 3 / November, 2004.

Hotho, A., J¨aschke, R.,, Schmitz1, C., Stumme, G. "Information Reterival in Folksonomies: Search and Ranking", in LNCS 4011, pp.411-426, 2006.

Huang, Y.C., Hung, C.C., Hsu, J.Y.: "You Are What You Tag", in AAAI, 2008.

Ioannidis JPA (2008) Measuring Co-Authorship and Networking-Adjusted Scientific Impact. PLoS ONE 3(7): e2778. doi:10.1371/journal.pone.0002778.

Kleinberg, J., "Analyzing the Scientific Literature in its online Context". Nature, in Web Focus on Access to the Literature, April, 2004

MacGarvie, M., "The determinants of international knowledge diffusion as measured by patent citations", in Econ. Lett. 87, pp. 121–126, 2005.

Marlow, C., Naaman, M., Boyd, M., Davis, M., "HT06, Tagging paper, Taxonomy, Flickr, Academic article, to read", in proceeding of the 17th conference on hypertext and hypermedia, in HT, Odense, Denmark, 2006.

Maurseth, P. B., and Verspagen, B.: "Knowledge Spillovers in Europe: A Patent Citations Analysis" in. Scandinavian Journal of Economics, Vol. 104, No. 4, pp. 531-545, 2002 Available at SSRN:http://ssrn.com/abstract=371854.

Michlmayr, E., Cayzer, S.: "Learning User Profiles from Tagging Data and Leveraging them for Personalized Information Access", in WWW Banff, Canada, 2007.

Mika, P.,:"Ontologies Are Us: A Unified Model of Social Networks and Semantics". In Proc. of 4th Intl. Semantic Web Conference (ISWC2005), 2005.

Park, G., Park, Y., "On the measurement of patent stock as knowledge indicators", in Technol Forecast Soc Change 73 (7), pp. 793–812, 2006.

Puntschart, I. and Tochtermann, K., "Online-Communities and the "un"-importance of e-Moderators", in Proceedings of Networked Learning 2006, Lancaster (UK), April 2006

Scharnhorst, A., Wouters, P. "Web Indicators – a new Generation of S & T Indicators", in international journal of scientometrics, Informmetrics and Bibliometrics, Vol. 10, issue 1, 2006.

Sorenson, O. and Singh, J., "Science, Social Networks and Spillovers" (December 26, 2006). Available at SSRN: http://ssrn.com/abstract=953731.

Tsai, W. "Knowledge Transfer in Intra-Organizational Networks: Effects of Network Position and Absorptive Capacity on Business Unit Innovation and Performance", in Academy of Management Journal, 44(5), 996-1004, 2001.

Us Saeed, A., Stocker, A., Hoefler, P., Tochtermann, K., "Learning with the Web 2.0: The Encyclopedia of Life", in Conference ICL2007, Villach, Austria, 2007.

Us Saeed. A, Afzal, M.T.,Latif, A., Stocker, A., Tochtermann, K., Does Tagging indicate Knowledge diffusion? An exploratory case study, In Proc. of 3rd ICCIT pp.605-610 , 2008a.

Us Saeed, A. Afzal, M.T. Latif, A. Tochtermann, K., Citation rank prediction based on bookmark counts: Exploratory case study of WWW'06 papers, INMIC 2008. IEEE International pp. 392 - 397, Dec. 2008b.

Wu, H., Zubair, M., Maly, K.: "Harvesting Socail Knowledge from Folksonomies", in HT, Odense Denmark, 2006.