Thematic Digital Libraries vs. Wikipedia's "One Size Fits All" - Lessons Learned

Barthel, Simon and Balke, Wolf-Tilo

Abstract— Today the Web serves as the central information hub for almost all areas of daily life, where issuing a simple keyword query offers access to a dazzling array of information. And this does not only cover trivial information, also in the area of professional or scientific information there is a clear trend towards easily accessible information provisioning via the Web, e.g. in digital libraries, open access journals, or topically focused forums or newsgroups. But here, querying mostly is much more complex and offers a vast variety of community-specific interfaces, specific indexing schemes, and metadata-based access. This increased complexity leads to the question whether such effort is really needed or if general purpose knowledge portals, like for instance Wikipedia, would already be sufficient even for sophisticated tasks with a clear thematic focus. In this paper we explore the challenges and chances of specialized thematic digital libraries reviewing typical use cases from different disciplines like chemistry or mathematics and argue that although one size does not fit all, there is a lot to learn from general purpose portals.

Index Terms— *digital libraries, metadata generation, metadata indexing, Wikipedia*

1. INTRODUCTION

LIBRARIES have been in existence for over 5000 years: from the oldest known collections of cuneiform written on clay tables in Sumer about 3000BC via the great library of Alexandria, where for the first time the idea of collecting all scholarly knowledge of the world in a single place was implemented, to today's central national libraries like the US Library of Congress containing more than 32 million books, 61 million manuscripts as well as millions of newspapers, scholarly articles, microfilm reels, comic books, and maps in more than 400 languages. With growing collections it became obvious that for supporting advanced search tasks and for managing the heterogeneity in this amount of data sophisticated schemes for bibliographic indexing and topical annotations were needed. Prime examples are the Dewey Decimal Classification (DDC) or the Library of Congress Classification (LCC). But already at an early stage there was a distinction between generally applicable *universal schemes* like DDC and LCC and more expressive *specific classification schemes* for particular subjects or types of materials like for instance the National Library of Medicine classification system (NLM).

With the increasing trend of digitization for improved accessibility the basic ways of describing library items slowly began to change. On one hand increasingly powerful computer systems enabled a self-description of documents even in large corpora (e.g., full text search, information retrieval, and text mining) as opposed to expensive manual indexing. On the other hand after the advent of the World Wide Web the active interlinking of documents and their subject-specific compilation in focused Web-accessible collections enabled powerful application scenarios like shared research environments or workplaces.

As far as information search is concerned both kinds of basic indexing schemes left their mark in digital collections and knowledge portals. Search engine technology of the *universal type* uses simple keyword search with only a modicum of general classification information like e.g. in Wikipedia. Navigational or facetted interfaces as often used in thematic digital libraries use *specific classifications* like subject-heading systems, taxonomies, thesauri, or any other type of structured controlled vocabulary.

But in contrast to the obvious advantages of higher expressiveness in querying using specific classification schemes in physical collections, where search was necessarily only based on metadata, the advantage of specific classification schemes in digital collections, where text-based indexing is easy and classification based on machine learning is possible, is an active area of discussion. Especially the growing heterogeneity of user communities due to increased interdisciplinarity and the rising costs for the classification schemes' maintenance may be good points in favor of abandoning subject specific classifications leading to less complex information access. The exponentially growing amount of information to be searched on the other hand may call more than ever for subject-specific methods for filtering and retrieval, where ease of access is sacrificed for search effectiveness.

In this paper we present lessons learned from designing search interfaces for thematic digital libraries in different fields like chemistry, mathematics, and medicine.

Manuscript received April 14, 2014.

S. Barthel and W.-T. Balke (contact author) are with the Institute for Information Systems (IfIS), University of Braunschweig, Germany (e-mail: { barthel, balke}@ifis.cs.tu-bs.de).

2. CHALLENGES FOR TODAY'S DIGITAL LIBRARIES

Since the 70s lots of research has been done in the area of Information Retrieval and Machine Learning and the continuous improvement of processing power and networking capacities smoothed the way for big search engines like Google, Bing, Yahoo, etc. to become the first choice for casual users who wants to satisfy their information need.

Still, for scholarly users there are some specialized products like Microsoft Academic Search, Google Scholar or CiteSeerX, which can be used for literature search in arbitrary domains with acceptable quality of results. In contrast to general search engines here the document base is more restricted and some advanced features like author search or document ranking by numbers of citations are supported. In fact, today these products are the biggest competitors for digital libraries. Indeed it seems most natural to use e.g. Google Scholar for scientific searches, if one uses Google Web Search for everyday queries. If the results are satisfying, there is no reason why researchers should bother to constantly switch between several focused digital libraries and manually integrate information for a hypothetical improvement of search results. So the question is, what exactly is the benefit of a digital library?

A number of international initiatives are already building vast digital archives of global content for universal access. For encyclopaedic knowledge Wikipedia is the major resource worldwide offered in a vast variety of languages. For cultural heritage the two largest initiatives today are the Open Content Alliance and the Europeana, both of which already provide collections of several million curated and quite well-maintained items. However, today both are basically huge databases storing digital items and allowing the retrieval based on metadata. In particular, these systems only offer unitary navigational access to different disperse segments of information, although providing links to other related content. This makes the process of building a holistic view and a deep understanding of some topic (or information need) difficult and time consuming.

In contrast digital libraries spend enormous efforts on manually curating their collection and enabling customer-centered access. Every item is carefully indexed and enriched with metadata - e.g. structural elements like subject headings are annotated, authors or publishers are identified, semantic keywords are assigned and much more. The results of a query can then for example be filtered by the annotated metadata using a facetted search. Another aspect that distinguishes general purpose search engines from digital libraries is the quality management regarding the indexed items. For digital libraries it is important that every item corresponds to a real publication produced by a real publisher or by a real author while a search engine just crawls and indexes everything of potential relevance that appears to be a scientific paper or a book.

This quality management becomes more and more difficult. Consider for instance TIB Hannover, the German National Library of Science and Technology: currently about 90,000 metadata entries have to be manually annotated each day with strongly increasing tendency. Moreover, there are not only classical publications to be indexed, but also a growing amount of primary research data in heterogeneous formats ranging from experimental data sets via simulation data to descriptive models. However, as libraries have to annotate more and more data with the same amount of financial resources, the problem of retaining the high quality of annotations becomes more pressing from year to year. One way to cope with this problem is to focus closely on one domain and build dedicated subject-oriented digital library where the range of metadata values is limited.

The benefit of such a thematically narrow library is not only that due to the special focus less data items have to be catered for, but also that it can be more responsive to the special requirements of the domain in the sense of value-added services. For example depending on the user's expertise information can be provided in different abstraction levels or search interfaces can be tailored with respect to the domain.

The problem that the information need of arbitrary users cannot be satisfied by one source can be well illustrated with an article about the Higgs Boson published by Scientific American [1], which caters to quite a heterogeneous readership. The magazine tried to answer the question of the nature of the Higgs Boson in three different ways:

- With an introductory text understandable for the general reader featuring 2 paragraphs length provided by an author from the Northeastern University
- With a more fundamental text of 6 paragraphs provided by an author from the Santa Cruz Institute for Particle Physics
- With a text aiming at experts in the respective domain over 9 paragraphs provided by an author of the Fermi National Acceleration Laboratory.

Clearly, performing such an amount of journalistic editing to satisfy the information need of every possible target user group for all topics cannot be the default case. A stricter focus on the other hand combined with domain specific personalization techniques as provided by most subject-centered digital libraries has great potential to provide users with the exact level of information that matches their information need and expertise.

Also, when performing research in special domains there are lots of search requirements and access patterns that only apply to that domain like e.g. querying models, experimental raw data,

test corpora or structured domain knowledge. But whenever a search interface for a domain independent library is provided, there are two extreme choices: either all kinds of access can be supported whether they are applicable to all items or not, or the interface can only offer access using common entities that are available in every scientific domains, like e.g. authors, articles, or books. In the first case the interface becomes cluttered up to the stage of being unusable, in the second case search expressiveness is severely hampered. Thus, most interfaces aim at a more or less satisfying compromise.

In the domain of chemistry researchers might for instance be interested in an interface that understands a chemical formula, so that they can perform a query like C_6H_6. They will then certainly find out that C_6H_6 corresponds to the molecule benzol – and 217 other molecules that all have 6 hydrogen and 6 carbon atoms but are structured in different ways. The important question that has then to be answered by the retrieval system is: what molecule is most relevant to the researcher? Since molecules are not per se relevant or irrelevant, this question is impossible to answer until the chemist's research context or search context is known. Also, this search context and the entities in the database have to be represented semantically to determine how strong an entity matches with a search context. This is normally done by using semantic metadata.

Semantic metadata, in the form of structured domain knowledge (like e.g. ontologies, taxonomies or controlled vocabularies) exists for all kinds of scientific domains, like the Mathematics Subject Classification (MSC) for mathematics, the Open Biomedical Ontologies (OBO) for biology or the Medical Subject Headings (MeSH) for medicine. Once the entities of interest are annotated with the respective domain knowledge with high quality and a search context is also represented using the same form of structured domain knowledge, it is possible to rank entities with respect to the user's context.

In the chemical domain the task of annotating such semantic metadata is e.g. performed by the Chemical Abstracts Service (CAS) with a huge amount of effort. The database containing these annotations is however not freely accessible but a standard CAS user license costs around 30,000 USD. For researchers, the purchase of such a user license can of course be a burden. Fortunately, not only commercial services provide semantic annotations but also digital libraries like e.g. the Zentralblatt MATH (zbMATH) which is the biggest digital library for mathematics in Europe. Every article being published in the zbMATH is annotated with respect to the MSC that is also maintained by the zbMATH in cooperation with Mathematical Reviews. However, the increasing amount of digitally available data becomes a continuously growing problem for digital libraries as more data needs to be annotated with the same amount of financial resources.

A solution to this problem might be to use automated text categorization approaches to automate the process of indexing. An extensive large scale study of state-of-the-art machine learning technologies applied on the corpus of the Zentralblatt MATH has already been conducted [2]. The authors examined the text contained in the title and the abstract of the mathematical articles and also performed advanced methods to prepare the formulae contained in the abstracts for automated classification. Figure 1 shows the confusion matrix of the resulting classifiers for the 63 top level classes of the MSC. The experiments show that only few top level MSC categories can be annotated with acceptable quality while most categories have a high confusion among each other. Consequently, the experiments trying to specify the definite MSC class on the lowest level of the MSC performed even worse.

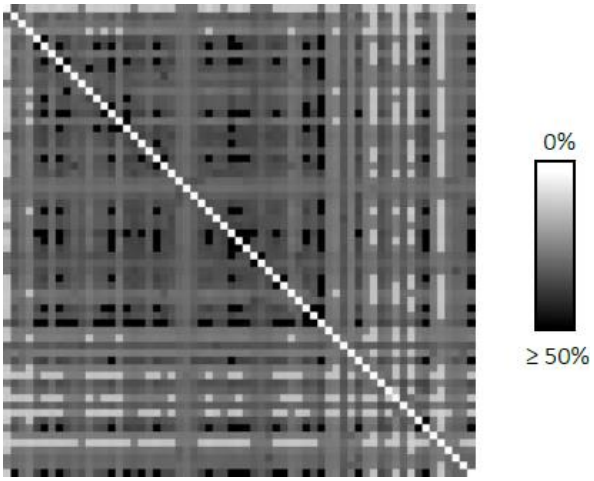

Figure 1: Confusion between the 63 Top Level classes of the MSC using the classifiers in [2]

Of course the problem that manually created taxonomies that have been manually maintained over centuries cannot be annotated automatically out-of-the-box with high precision is not a problem that is restricted to mathematics. This indicates that digital libraries will very soon not be able to retain the high quality semantic annotations as manual indexing becomes more and more infeasible and automatic indexing is too inaccurate. It therefore seems that digital libraries will in future be forced to either rely on low quality automatic annotations or to restrict the semantic annotations to the most important journals.

In [3] the authors present an approach that enabled context searches for chemical entities, despite the documents' lack of suitable context annotations for the domain. The authors presented a similarity measure using cross-domain knowledge gathered from Wikipedia. Even though the quality of their annotations was not overwhelming, the annotations were still useful to specify and personalize a user's search context. For digital libraries this means that low quality

annotations are still better than no annotations at all to provide personalized context sensitive information retrieval. However, it is questionable if this matches with the high demands of quality that are associated with digital libraries.

3. *Focused Digital Libraries*

It stands to reason that sooner or later digital libraries need the assistance of machine learning approaches to cope with the growing amount of digitally available information. Unfortunately, all evidences show that automating the annotation work of the editors in digital libraries will not work out-of-the-box. The main reason for this problem is that machine learning algorithms were first regarded at the very end of a long evolution of semantic metadata annotation and were not regarded from the start.

To cope with the growing flood of newly published documents and still be able to scale, the focus of a subject-oriented digital library must be to phase the structured domain knowledge in form of ontologies, taxonomies or controlled vocabularies with machine leaning technologies. This means that structured domain knowledge must be designed in a way that it can be annotated using machine leaning technologies but still domain experts must be able to understand the semantics of the automatically generated annotations. Finding and maintaining such structured domain knowledge is certainly not an easy task but it is a task that does not become harder with increasing amount of documents.

Below, we will give a short introduction in related research on topical information extraction as well as an overview of current trends in that topic.

3.1. *Subject-centered information extraction*

Every scientific domain has developed a certain dialect where natural language is mixed with domain specific terminology. The ability to understand and extract this terminology reveals great opportunities for the creation of value-added services for subject-centered retrieval systems.

In the area of mathematic, for instance it is natural to use formulae in running text. The formulae in mathematical texts are used as arbitrary parts of speech like adjectives, objects, subjects or whole sub sentences. The authors of [4] e.g. describe the contribution of their work as follows:

> "[…] We prove in particular that if $f(x) = ax^n + x^m$ permutes $\mathbb{F}_q$, where $n > m > 0$ and $a \in \mathbb{F}_p^*$, then $p - 1 \leq (d-1)d$, where $d = \gcd(n-m, p-1)$, and that this bound of p, in terms of d only, is sharp. […]"

We see here that the formula $f(x)$ is used as a subject and $\mathbb{F}_q$ as an object in the first sentence. The authors further state that "if $f(x)$ permutes $\mathbb{F}_q$, then $p - 1 \leq (d-1)d$" using the inequation $p - 1 \leq (d-1)d$ as a conditional sub sentence. The same happens with the term $n > m > 0$ to express that n is greater than m and both n and m are greater than 0. It is also striking that strong domain specific notations were used when constraining the range of certain variables with a sub sentence starting with a leading "where". A parser that understands formulae together with the context around the formulae would also be able to distinguish between an article that *uses* a certain formula or *proves* a certain formula or under what *condition* an article proves a certain formula.

In [5] the authors propose a graph based approach for mathematical knowledge management. The proposed theory graphs approach as a representation paradigm for mathematical knowledge that allowed to make the modular and highly networked structure of mathematics explicit and therefore machine-actionable. Using this approach for example in digital mathematic libraries, it reveals the potential for computer-supported or even automatic representation, cataloging, retrieval, refactoring, plausibilization, and in some cases even application of mathematical knowledge.

To give another example let's look into the field of chemistry. If an organic chemist describes a synthesis procedure it may read as follows [6]:

> **"[…] 5-Cyclobutyl-2,3-dihydro-[1H]-2-benzazepine 82:** Potassium carbonate (0.63g, 4.56mmol) and thiophenol (0.19g, 1.69mmol) were added to the 2-nitrobenzene sulfonamide **50** (0.50g, 1.302mmol) in N, N-dimethylformamide (33cm^3) at room temperature and the mixture was stirred for 16h. Deionised water (50cm^3) was added and the aqueous phase was extracted with ethyl acetate (5×50cm^3). The organic extracts were dried ($MgSO_4$) and concentrated under reduced pressure to give the title compound **82** (0.259g, 1.302mmol, ca. 100%) as an oil used without further purification. […]"

Again, very strong domain specific terminology and conventions can be observed. Procedure descriptions in the field of organic chemistry always show a high amount of domain specific terminology (like X was added to Y at room temperature, X was extracted with Y, etc.). A system that is able to transform above example into a structured form would also allow to build a system to compare and search for synthesis procedures. Of course, this is only a very specific example - an overview summarizing state of the art methods for information extraction used in chemistry is described in [7].

These two examples illustrate what benefits domain specific natural language processing can provide compared to general purpose approaches. Both examples give much opportunity for the application and development of automated information extraction methods like e.g. for the extraction of named entities, domain specific synonym detection or disambiguation. Scientific papers are also a promising target for the extraction or the computer-supported generation of structured knowledge, like ontologies, taxonomies or thesauri.

Currently, structured domain knowledge is almost exclusively created completely manually. Having in mind that this structured knowledge must cover all aspects in the current research in the domain, is becomes clear that the creation and maintenance of such knowledge is related to a huge amount of effort and high personnel costs. Additionally, to ensure completeness, soundness and actuality is certainly not an easy task. Therefore, research trying to generate domain knowledge automatically or at least semi-automatically, became a popular area of research in computer science. The general idea to generate domain knowledge is to extract Salient Terms and relationships out of the document corpus of interest and represent them as ontological knowledge. Basic technologies for this task range from Natural Language Processing [8], [9] over probabilistic language models [10], [11] to statistical approaches adapted from the area of Information Retrieval [12]. In recent years, research has also applied on so called Folksonomies [13] where Data Mining Techniques are applied on metadata provided by users (like e.g. user tags). These so called "Light-Weight-Ontologies" can for example be used as a compromise between carefully created ontological knowledge and a loose collections of metadata.

Another point to consider is what sources should be regarded for the annotation of metadata. Of course, the obvious source is the text belonging to the document itself but the whole picture of a document's context is naturally not restricted to the document's text. Popular sources to enrich the amount of information for a document are citation or author networks. The properties of citation networks are known very well and have been studied over many years [14], [15]. Citation/author networks can for instance be used to boost the quality of annotations [16] or to rank publications or authors based on the structure of the citation graph [17], [18].

Another popular, newly occurring source of metadata is the Social web. Recently, analyzing this data to estimate the impact and quality of scholarly publications gets more and more popular under the term of altmetrics [19], [20]. These metrics try to reflect activity in social media services with the purpose of gathering scholarly impact besides the traditional citation based metrics. Tracking these activities, it is possible to monitor the manner in which scholarly documents are disseminated and discussed in a narrow time frame [21]. The role of social media in scholarly communication has been investigated in several studies including their use in dissemination [22], conference chatter [19], science popularization [23], and promotion of scholarly products [24]. In addition, several tools have been introduced to facilitate the use of Altmetrics, e.g. PlumX, ImpactStory, Altmetrics and Scholarometer [25].

As the target of this article is only to give an intuition of the capability of topical digital libraries, we will not go into details of information extraction here. For more information on this topic we therefore refer to [26].

3.2. *Creating Useful Rankings*

The ultimate goal of a digital library (as for every information retrieval system) is to answer the users' information needs with plausible rankings. As already stated, to succeed in this task domain specific information extraction is required to represent the knowledge and semantics of a domain and to annotate documents and entities in the domain, respectively. The basis of building such a ranking is a similarity measure.

However, computing a similarity of two entities based on given sematic annotation is also not an easy task. In chemistry there exist a whole stack of similarity measures between substances in use, like the Tanimoto measure, the Russel-Rao dissimilarity, the Yule distance and much more. The question what similarity measure is best suited for chemists is not easy to answer. In [27] the authors compared and evaluated different similarity measures and tried to find out if the metrics are redundant or if they represent different concepts of similarity. An important observation was that that the similarity measures show almost no correlation, meaning that they certainly represent different aspects or definitions similarity. It was also shown that the assessment of the quality of a similarity measure is highly dependent on the individual chemist, suggesting that different similarity measures correspond to different ways of chemists' perception of similarity. The authors also showed that by using feedback provided by the users the quality of rankings could be improved significantly after only very few feedback cycles. This observation shows that subject-oriented libraries can improve the search quality greatly by using domain specific structured knowledge. For text search this improvement might still be a nice-to-have feature, but for domain specific entities like e.g. molecules the annotation of domain specific structural knowledge is mandatory to provide a plausible ranking.

4. CONCLUSION

This article aims to raise the awareness of the necessity of subject-oriented digital libraries as opposed to "one size fits all"-style knowledge repositories. As textual structures, conventions, structured knowledge as well as requirements to a retrieval system differ tremendously among different domains, subject-oriented digital libraries can produce great benefit for experts in the respective domains.

An essential task to provide domain specific value-added services is the annotation of structured domain specific knowledge like e.g. keywords from ontologies, taxonomies or controlled vocabularies. However, the annotation of these keywords as is it performed today will not work

out in the future. Also, automated indexing approaches where state-of-the-art machine learning algorithms are “plugged” on top of a long evolution of manual indexing work are proven to not work properly. We therefore claim that in order to retain high quality semantic annotations and on the other hand still be able to scale, it is necessary to develop structured domain knowledge that sufficiently reflects the semantic of the domain but is able to be annotated automatically by using state-of-the-art machine learning technology.

We also gave an insight what topical digital libraries can target. We showed several examples of domain specific texts and discussed what kind of information extraction methods can be applied to the texts that are restricted to that particular domain. We also looked into different sources to obtain metadata, especially we introduced the newly occurring idea of altmetrics that utilizes data from the social web to assess the impact or quality of a publication.

References

[1] “What exactly is the Higgs boson? Have physicists proved that it really exists?,” *Scientific American*. [Online]. Available: http://www.scientificamerican.com/article/what-exactly-is-the-higgs/. [Accessed: 16-Apr-2014].

[2] S. Barthel, S. Tönnies, and W. Balke, “Large-Scale Experiments for Mathematical Document Classification,” *Int. Conf. Asia-Pacific Digit. Libr.*, vol. 15, pp. 83–92, 2013.

[3] B. Köhncke and W. Balke, “Context-Sensitive Ranking Using Cross-Domain Knowledge for Chemical Digital Libraries,” in *17th International Conference on Theory and Practice of Digital Libraries (TPDL)*, 2013, p. pp 285–296.

[4] M. Ayad, K. Belghaba, and O. Kihel, “On permutation binomials over finite fields,” *Bull. Aust. Math. Soc.*, vol. 89, no. 01, pp. 112–124, Mar. 2013.

[5] M. Kohlhase, “Mathematical Knowledge Management: Transcending the One-Brain-barrier with Theory Graphs,” *EMS Newsletter*, 2014.

[6] B. Bradshaw, P. Evans, J. Fletcher, A. T. L. Lee, P. G. Mwashimba, D. Oehlrich, E. J. Thomas, R. H. Davies, B. C. P. Allen, K. J. Broadley, A. Hamrouni, and C. Escargueil, “Synthesis of 5-hydroxy-2,3,4,5-tetrahydro-[1H]-2-benzazepin-4-ones: selective antagonists of muscarinic (M3) receptors.,” *Org. Biomol. Chem.*, vol. 6, no. 12, pp. 2138–57, Jun. 2008.

[7] L. Hawizy, D. M. Jessop, N. Adams, and P. Murray-Rust, “ChemicalTagger: A tool for semantic text-mining in chemistry.,” *J. Cheminform.*, vol. 3, p. 17, 2011.

[8] M. A. Hearst, “Automatic Acquisition of Hyponyms from Large Text Corpora,” in *Proceedings of the 14th International Conference on Computational Linguistics*, 1992, pp. 539–545.

[9] E. Stoica, M. Hearst, and M. Richardson, “Automating Creation of Hierarchical Faceted Metadata Structures,” *Proc. NAACL HLT 2007*, pp. 244–251, 2007.

[10] P. Cimiano, S. Handschuh, and S. Staab, “Towards the self-annotating web,” in *Proceedings of the 13th conference on World Wide Web - WWW ’04*, 2004, p. 462.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[12] M. Sanderson and B. Croft, “Deriving concept hierarchies from text,” in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’99*, 1999, pp. 206–213.

[13] R. Jäschke, A. Hotho, C. Schmitz, B. Ganter, and G. Stumme, “Discovering shared conceptualizations in folksonomies,” *Web Semant.*, vol. 6, pp. 38–53, 2008.

[14] S. Lehmann, B. Lautrup, and A. D. Jackson, “Citation networks in high energy physics.,” *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, vol. 68, p. 026113, 2003.

[15] M. Franceschet, “The large-scale structure of journal citation networks,” *J. Am. Soc. Inf. Sci. Technol.*, p. n/a, 2012.

[16] X. Li, H. Chen, Z. Zhang, and J. Li, “Automatic Patent Classification using Citation Network Information: An Experimental Study in Nanotechnology,” in *PROCEEDINGS OF THE 7TH ACM/IEE JOINT CONFERENCE ON DIGITAL LIBRARIES*, 2007, pp. 419–427.

[17] Y. Ding, E. Yan, A. Frazho, and J. Caverlee, “PageRank for ranking authors in co-citation networks,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 11, pp. 2229–2243, Nov. 2009.

[18] S. Maslov and S. Redner, “Promise and pitfalls of extending Google’s PageRank

algorithm to citation networks.," *J. Neurosci.*, vol. 28, no. 44, pp. 11103–11105, 2008.

[19] K. Weller, E. Dröge, and C. Puschmann, "Citation Analysis in Twitter : Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences," in *1st Workshop on Making Sense of Microposts*, 2011, pp. 1–12.

[20] J. Priem, D. Taraborelli, P. Groth, and C. Neylon, "altmetrics: a manifesto – altmetrics.org." [Online]. Available: http://altmetrics.org/manifesto/. [Accessed: 18-Mar-2014].

[21] X. Li, M. Thelwall, and D. Giustini, "Validating online reference managers for scholarly impact measurement," *Scientometrics*, vol. 91. pp. 461–471, 2012.

[22] E. S. Darling, D. Shiffman, I. M. Côté, and J. a. Drew, "The role of Twitter in the life cycle of a scientific publication," *PeerJ Prepr.*, vol. 1, p. 16, 2013.

[23] C. R. Ugimoto and M. Thelwall, "Scholars on Soap Boxes: Science Communication and Dissemination via TED Videos," *J. Am. Soc. Inf. Sci. Technol.*, 2012.

[24] B. Cronin, "Metrics à la mode," *J. Am. Soc. Inf. Sci. Technol.*, vol. 64, no. 6, pp. 1091–1091, Jun. 2013.

[25] J. Kaur, D. T. Hoang, X. Sun, L. Possamai, M. JafariAsbagh, S. Patil, and F. Menczer, "Scholarometer: A Social Framework for Analyzing Impact across Disciplines," *PLoS One*, vol. 7, 2012.

[26] W.-T. Balke, "Introduction to Information Extraction: Basic Notions and Current Trends," *Datenbank-Spektrum*, vol. 12, no. 2, pp. 81–88, May 2012.

[27] S. Tönnies, B. Köhncke, and W.-T. Balke, "Taking Chemistry to the Task – Personalized Queries for Chemical Digital Libraries," in *Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2011, pp. 325–333.