ARTICLE IN PRESS

Information Processing and Management xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks

B. Berendt [a,*], B. Krause [b], S. Kolbe-Nusser [c]

[a] Dept. of Computer Science, K.U. Leuven, Leuven, Belgium
[b] Inst. of Knowledge and Data Engineering, University of Kassel, Germany
[c] Inst. of Information Systems, Humboldt University Berlin, Berlin, Germany

ARTICLE INFO

Article history:
Received 16 September 2008
Received in revised form 28 April 2009
Accepted 6 August 2009
Available online xxxx

Keywords:
[H.2.8] Database management – database applications – data mining
[H.3.7] Information storage and retrieval – digital libraries – user issues
[H.3.3] Information storage and retrieval – information search and retrieval – search process, information filtering
[H.3.5] Information storage and retrieval – online information services – Web-based services
[K.3.2] Computers and education – computer and information science education – literacy
Citation analysis

ABSTRACT

Many powerful methods and tools exist for extracting meaning from scientific publications, their texts, and their citation links. However, existing proposals often neglect a fundamental aspect of learning: that understanding and learning require an active and constructive exploration of a domain. In this paper, we describe a new method and a tool that use data mining and interactivity to turn the typical search and retrieve dialogue, in which the user asks questions and a system gives answers, into a dialogue that also involves sense-making, in which the user has to become active by constructing a bibliography and a domain model of the search term(s). This model starts from an automatically generated and annotated clustering solution that is iteratively modified by users. The tool is part of an integrated authoring system covering all phases from search through reading and sense-making to writing. Two evaluation studies demonstrate the usability of this interactive and constructive approach, and they show that clusters and groups represent identifiable sub-topics.

1. Introduction

Bibliometric analyses and other information retrieval/data mining (IR/DM) techniques are powerful instruments for unlocking the contents of scientific publications. Via the Web, these techniques can be made available to everyone, and indeed an increasing number of Digital Libraries and search engines offer citation-network analysis functionality. However, existing proposals often neglect a fundamental aspect of learning: that understanding and learning require an *active* and *constructive* exploration of a domain.

In this paper, we describe a new method and a tool that use machine intelligence and interactivity to turn the typical *search and retrieve* dialogue, in which the user asks questions and a system gives answers, into a dialogue that also involves

* Corresponding author.
E-mail address: Bettina.Berendt@cs.kuleuven.be (B. Berendt).
URLs: http://www.cs.kuleuven.be/~berendt (B. Berendt), http://www.kde.cs.uni-kassel.de/krause (B. Krause), http://www.defaultdomain.de (S. Kolbe-Nusser).

0306-4573/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2009.08.002

sense-making, in which the user has to become active by constructing a bibliography and a domain model of the search term(s). We call this activity *context creation* (for a search and search term).

This tool is part of an integrated authoring system covering all phases from search through reading and sense-making to writing. The goal of this system is to show scientists what they can gain from (high-quality) citation data, in order to motivate them, *in the same environment*, to contribute such high-quality data – again using techniques of IR/DM (Berendt, Dingel, & Hanser, 2006). The goal is to create a positive-feedback loop in which more and more correct metadata are created and used for scholarly progress. To the best of our knowledge, no other comparable tool or service exists that is modelled on the whole process of scientific writing and that accompanies authors in their standard environment.

After a survey of related work in Section 2 of this paper, Section 3 describes our system's architecture and the algorithmic methods and usage interface of the tool. In Section 4, we summarise two evaluation studies that assess qualitative and quantitative quality criteria. Section 5 concludes with an outlook.

2. Motivation and related work

A central goal of our tool is to support an active construction of a search-context and domain structure by the user, based on pertinent information in the documents. The measurement of such pertinent information has been studied by the discipline of bibliometrics for several decades, and two types of information have emerged as central in these investigations: citations and texts.

Citation analysis serves to detect the structure and evolution of science: generic vs. specialised authors and topics, "specialty narratives", the changing "frontiers of science", and changes in paradigms (see Chen, 2003 for an overview).

Tools to tap this huge potential of citation data are therefore interface elements for repositories like http://citeseer.ist.psu.edu/ http:/citeseerx.ist.psu.edu www.citebase.org, www.slac.stanford.edu/spires/hep, portal.acm.org, scholar. google.com, www.arxiv.org or http://www.thomsonreuters.com/products_services/scientific/Web_of_Science. All repositories offer some form of *topical* and *metadata* search. *Navigation* in the document network is generally based on lists of documents citing or cited-by the currently-viewed document. Links to similar documents (based on text, co-citation, or bibliographic coupling) are offered in some repositories. Recommendations based on such local similarities have been found to be useful in suggesting new references (McNee et al., 2002).

Various tools show *bibliometric measures* like number of citations, rankings of the authors or journals of the found publications, etc., or *visualisations* of the repositories and their semantics (Chen, 1999; Chen, 2003).

A problem of all these tools is that they do not support the user in constructing a domain model of a researched field. They highlight certain aspects (like key articles that caused a change in the field's discourse, cf. Chen, 2006) and focus on the presentation of an automatically generated model, or they help in navigating but do not support model building. Also, tools that perform overview analyses on large datasets (Small, 1994; Chen & Carr, 1999; Chen, 2003; Chen, 2006) are generally not available to the public, or they operate on pre-processed test sets rather than on live Digital Libraries (DLs).

Recently, analysis tools have been developed that are available either for local installation and use with arbitrary datasets on the client's computer, or available via the Web tapping a live DL. An example are the "analytic tools" at www.ist-world.org working on a publication database (their functionality is based on the textual-similarity plus latent semantic analysis (LSA) in Fortuna, Grobelnik, & Mladenic, 2005). Search engines like www.clusty.com or www.kartoo.com employ similar text-based techniques on Web resources. However, the user can neither choose the level of granularity nor interactively label clusters. OntoGen(http://ontogen.ijs.si Fortuna, Mladenic, & Grobelnik, 2006) does involve the user in the specification of the number of clusters and labels. It leverages textual similarity and allows the user to build a hierarchy (ontology) of topics by recursively clustering documents. A well-known precursor of this method was proposed by Cutting, Pedersen, Karger, and Tukey (1992). Feng, Jeusfeld, and Hoppenbrouwers (2005) envisage future DLs in which knowledge thus learned adds a "knowledge subspace" (which could be queried directly for factual information) to the usual "document subspace" (which can be queried for relevant documents). All these tools and visions depend solely on textual similarity and do not exploit the expressiveness of citation information. Also, they have no or very limited possibilities of improving on the automatically generated clusters (which do not always make sense).

Textual similarity including LSA has also been employed in bibliometrics (e.g., Janssens, Leta, Glänzel, & De Moor, 2006). Janssens, Glänzel, and De Moor (2008) gave an overview of studies published since the seminal work of Braam, Moed, and van Raan (1991), which have shown that hybrid clustering methods that incorporate text and citation information can outperform clustering methods that use only one of these components, and they proposed a new hybrid method that outperforms both. Another form of hybrid clustering combines text-based document similarities with author similarities; this can be used for example for expert finding (Tho, Hui, & Fong, 2007). These tools are not available for general use or for accessing live DLs, and they have no context-creation components.

Interactive sense-making requires tightly interrelated activities of searching and representation forming (such as context creation), possibly enhanced by other activities like writing. The importance of this holistic view is increasingly being recognised, cf. (Qu & Furnas, 2008; Twidale, Gruzd, & Nichols, 2008). Several reference management tools support sense-making by allowing users to manually create categories. The system of Zhang, Qu, Giles, and Song (2008) relies on the structure

Please cite this article in press as: Berendt, B., et al. Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.08.002

found in the ACM subject hierarchy placement of publications. Bier, Good, Popat, and Newberger (2004) describe an interface for detail-and-context visualizations of such a structure. However, none of these tools leverage machine intelligence.

In sum, many solutions for analysing publications or documents exist, but all of them, when viewed from the perspective of our goal of active and constructive domain-model building, are only partial solutions. We encourage interactive model construction and labelling (as in Fortuna et al., 2006), leverage citation and textual information (as in Braam et al., 1991; Chen, 2006), and build on word profiles for characterising the content of document clusters (as in Braam et al., 1991; Chen, 2006; Fortuna et al., 2006; Janssens et al., 2008). In addition, we extend these functionalities by helps for personal productivity (storage) and sharing with others. Finally, we offer access to a live Digital Library (like the citation-analysis options in CiteSeer, Citebase, Google Scholar etc., or like the text-analysis and co-authorship analyses deployed at www.ist-world.org), but in contrast to all existing approaches, we realize this in an architecture based on Web services that is thereby modular, configurable and easily extensible.

3. The context creation tool

3.1. Functionality and user interface

The context creation tool has two front-ends in order to help users group documents in their preferred working environment: Web browser and MS Word. The first is familiar to and preferred by all users especially for searching, retrieving, and discussing with others, and popular also for PDF reading with standard browser plugins. The second is a highly popular writing environment for many users.

In the context creation tool, the first user input is a search term. Outputs are generated in three stages: First, a bibliographic database is searched, and the matching items are returned (*search* & *retrieval*). Second, they can be clustered. Clustering has a default mode for non-expert users and configurable options for expert use. The user is encouraged to label and describe the groups, and to modify the automatically-derived grouping to both reflect and develop her perception of the scientific domain in terms of a topic structure (*sense-making*). To help labelling, the top 10 TF.IDF keywords of the cluster are presented below the cluster name. The user can request a re-computation of the keywords after she has changed cluster contents by deletions or insertions.

Fig. 1 shows an example screenshot. In a search for literature on "RFID", the user has already labelled the first group (as "Security/privacy"), but the second still holds its default name ("Group 2"). Hyperlinks enable the user to directly retrieve the full text of a document. Based on pilot tests, a familiar 'search-engine interface' was provided, but enhanced by non-standard functionality: specifying the number of desired results, extending a search result by more and/or different-search-term results, and deleting and moving result documents between different clusters.

Third, the user can save and reload results, include them in personal documents, and make them available to others (*discussion*), via the Web in a format amenable for further processing (XML+CSS), see Fig. 2 for an example.

The tool is deployed at http://www.cs.kuleuven.be/~berendt/CiteseerCluster.

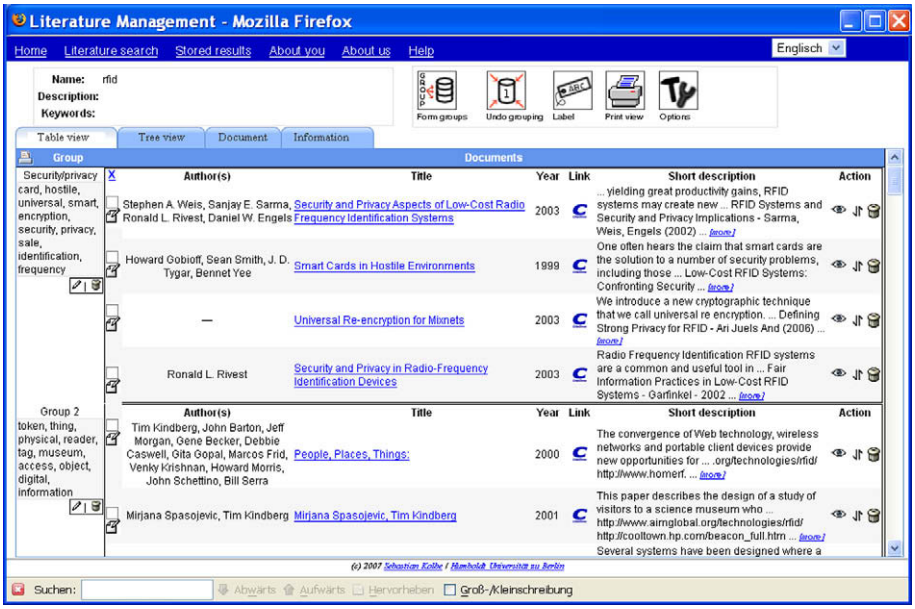

Fig. 1. Context creation: grouping and annotation.

Please cite this article in press as: Berendt, B., et al. Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.08.002

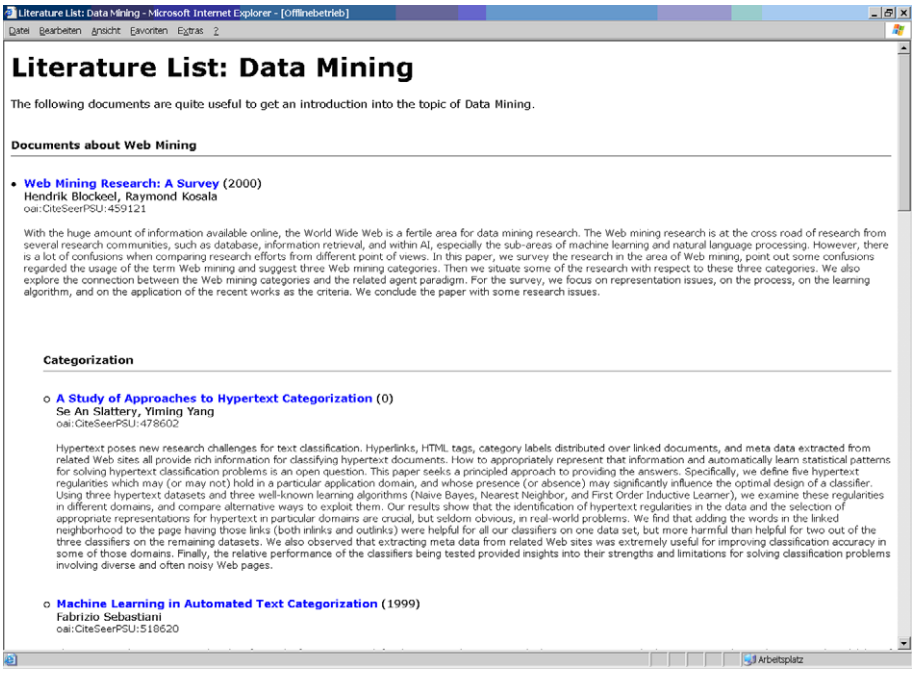

Fig. 2. Context creation: domain structure publishing for discussion.

3.2. Data processing and data sources

The user interfaces interact with a a back-end layer operating on a server (via Web server for the browser front-end and via Web services for the MS Word front-end). This layer accesses local data, search-engine Web services, and a remote repository as the main source of information on scientific literature. We use the CiteSeer repository because of its broad coverage[1] and rich structure, and also because it offers an OAI interface.[2]

Processing has five stages:

(1) The search term is transformed into a request to the Yahoo! Search Web service (http://developer.yahoo.com/search), with a restriction of results to those from the CiteSeer Web site.[3] A wrapper extracts the CiteSeer IDs from the returned URLs. The output of this step is a set of document IDs D' that are relevant to the search term, with the number of retrieved documents, $r' = |D'|$, governed by the user's input.
(2) For each document d in the result set, two lists of IDs are compiled: all documents that cite d, and all documents that are cited by d. This information is created from the citation matrix given by a local partial copy of the CiteSeer database. This local copy may not contain the most recent entries of the repository; thus steps (2)ff. in fact operate on $D \subseteq D'$ with $r \leqslant r'$ elements.[4]
(3) For each document $d \in D$, bibliographic metadata for result presentation (author, title, etc.) are retrieved from the local database.
(4) A similarity matrix for D is constructed.

Various *sources of similarity* are considered: co-citation, bibliographic coupling, and text-based similarities. Citation measures are derived from the citation matrix; text measures are derived from the abstract (a traditional choice, see e.g. Braam et al., 1991; Janssens et al., 2008). Text-based similarity is either based on words alone (optionally weighted by TF.IDF) or on LSA.

[1] CiteSeer focusses on computer science literature; corresponding sources for different academic fields should be investigated in the future.
[2] OAI is the Open Archives Initiative (www.openarchives.org), which has developed the OAI Metadata Schema and the OAI Metadata Harvesting Protocol for harvesting metadata about resources residing in separate repositories. A repository with an OAI interface allows programs to query and access the repository's metadata according to this protocol.
[3] This was preferred to search via CiteSeer (done in a previous version, see Berendt et al., 2006) because the amount of HTML parsing can be reduced (CiteSeer offers no Web service or OAI interface for searching); search was found to be more stable and result lists more comprehensive. Also, this detour leverages Yahoo!'s ranking to select a good initial result set.
[4] The earlier system version obtained citation relations from CiteSeer via the OAI interface (tag <oai_CiteSeer:relation type="References">). While this guarantees up-to-date information, it proved to be too slow and was therefore discontinued. The same holds for the subsequent step (3).

Please cite this article in press as: Berendt, B., et al. Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.08.002

The default is bibliographic coupling, which, like co-citation, is a classic type of document linkage extensively researched in bibliometrics. Co-citation has been found to be an excellent indicator of document similarity and also of global changes in an academic field (Small, 1973; Chen, 2003). Co-citation is limited by its focus on "pivotal documents" that have existed for long enough to be cited by many sources. Bibliographic coupling (Kessler, 1963) can help to group especially new documents faster: A new publication will associate itself to an existing cluster of similar documents by referring to the same sources. Both methods aggregate independent opinions about which documents are related, thus taking advantage of the robustness of 'collective intelligence'.

The user can choose between different *similarity measures*. For each choice of similarity source, the Jaccard coefficient, the Dice coefficient, or the cosine similarity can be used (see for example Tan, Steinbach, & Kumar, 2005 for definitions). To combine bibliographic coupling and co-citation information, the user can choose the Amsler measure (Bichteler & Eaton, 1980). To combine citation and text information, the user can specify a weighting factor for a linear combination between the two similarity matrices.[5]

As default, we use the Jaccard coefficient. The Jaccard coefficient is a popular, proven, and scalable method of measuring similarity between Web documents (Haveliwala, Gionis, Klein, & Indyk, 2002), and it has been used in co-citation (Small & Greenlee, 1980) as well as bibliographic-coupling (Bani-Ahmad, Cakmak, Özsoyoglu, & Al-Hamdani, 2005) analyses.

The resulting similarity values are then derived from the combination of similarity source and measure. For example, the bibliographic coupling similarity between documents d_1, d_2 with the Jaccard coefficient is defined as

$$sim_{bc}(d_1, d_2) = \frac{|\text{doc.s cited by } d_1 \text{ and by } d_2|}{|\text{doc.s cited by } d_1 \cup \text{doc.s cited by } d_2|}.$$

Only documents that *can* contribute to the numerator are considered, operationalized as documents that appeared in the minimum of the publication years of d_1 and d_2 or earlier (analogously for co-citation).

D may contain 'isolated documents' (Small & Griffith, 1974) that are not co-cited with anything, or that do not co-cite with anything. This can be detected by all-zero columns or rows in the citation matrix; both the row and column are deleted such that a $c \times c$ similarity matrix, with $c \leqslant r$, remains.

(5) The non-isolated documents from D are clustered using the toolkit CLUTO (http://www.cs.umn.edu/~karypis/cluto).

Different *clustering methods* can be chosen. Their selection was the result of prior experiments with the methods implemented in CLUTO. The user of our system can choose between hierarchical agglomerative clustering with complete link or UPGMA, "direct clustering" (a method similar to k-means, with a global optimality criterion), and RPR (repeated bisections with a global optimality criterion). The default is RPR.

The *number of clusters* can be determined by the user. In this case, its value is set to $min(n, c-1)$, with n the number of clusters specified by the user. The minimum $c-1$ guarantees that there is at least one two-element cluster. Alternatively, an optimal number of clusters is determined by the highest Silhouette value in the interval between 2% and 15% of the number of documents (cf. Tan et al., 2005; Janssens et al., 2008).

Tests showed that the clustering computation step, even if repeated, is very fast (the computation of the similarity matrix requires most time in the processing of most search queries).

If present, two additional groups are shown: isolated documents and documents whose citation links could not be analyzed because they are not in the local database. This is done to avoid arbitrary assignments while respecting that citation-based clusters do not represent the entire relevant literature that covers a topic (Braam et al., 1991).

An example. To illustrate this process, we present the following fictitious and necessarily small example of the stages described above: The repository contains documents $d_1, \ldots, d_7$. The user specifies a query and requests seven documents. The search engine identifies documents $D = \{d_1, d_2, d_3, d_4, d_5, d_7\}$ as relevant to the query. Thus, $r' = 6$ (the result would be the same for any user-specified number $\geqslant 6$). Since the very recent d_7 is not in the local database, $r = 5$. The local database contains (only) the following citation relations (first document cites second document): $(d_1, d_5), (d_1, d_6), (d_2, d_1), (d_3, d_1), (d_4, d_3), (d_5, d_2), (d_5, d_3)$. (3) Bibliographic metadata for documents in D are retrieved. (4) To keep the example simple, we consider the exclusive use of the Jaccard coefficient for bibliographic coupling as the similarity measure. We also assume that both d_2 and d_3 have been published after d_1, and both d_4 and d_5 have been published after both d_3 and d_2. The data yield: $sim_{bc}(d_2, d_3) = 1$, $sim_{bc}(d_4, d_5) = \frac{1}{2}$, and $sim_{bc}(d_i, d_j) = 0$ for all other pairs from D. d_1 does not co-cite with anything and is therefore the only element in the set of isolated documents. In sum, this produces $c = 4$ and a 4*4 similarity matrix. (5) The user chooses hierarchical agglomerative clustering and desires to see seven clusters. The system then forms $min(7, c-1) = min(7, 3) = 3$ clusters. The clustering solution is $\{\{d_2, d_3\}, \{d_4\}, \{d_5\}\}$ plus the isolated-documents group $\{d_1\}$ and the not-in-local-database group $\{d_7\}$. (The result would be the same for any user-specified number of clusters $n \geqslant 3$, and the further settings of the clustering algorithm do not affect the result.) $n = 2$ or a system-optimised cluster number would assemble d_4 and d_5 into one cluster.

[5] The latter was used in (Janssens et al., 2008); it gave better results than the use of only citation or text information, but worse results than a combination of the matrices based on Fisher's inverse chi-square. For our system and data, these combinations are still under investigation.

Please cite this article in press as: Berendt, B., et al. Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.08.002

4. Evaluation

We evaluated the tool with a combination of data-mining and usability quality measures. The purpose was not to evaluate the clustering of scientific literature *per se*; this has been done for instance in (Braam et al., 1991; Chen, 2006; Vladutz & Cook, 1984; Bani-Ahmad et al., 2005). Rather, our focus was the usefulness of the clustering and interaction for end users.

4.1. Cluster quality

Concept and measures Context creation is a knowledge-discovery task: designed to find "valid, novel, potentially useful, and ultimately understandable patterns" in data (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996). Here, the patterns are document groups.

Validity would ideally be assessed by traditional measures of cluster validity; specifically external or relative measures (Halkidi, Batistakis, & Vazirgiannis, 2002a; Halkidi, Batistakis, & Vazirgiannis, 2002b). External measures compare a clustering solution with a pre-defined, "ground-truth" classification. Relative measures compare different clustering solutions (for example, with different numbers of clusters) by indexes that combine the goals of maximizing intra-cluster similarities and minimizing inter-cluster similarities.

A ground truth does not exist, in general, for scientific literature, and it cannot exist, *a forteriori*, for document sets defined by arbitrary search terms. An application of relative measures for sample datasets produced, due to the sparsity of the similarity matrices, "optimal" results with a large number of very small (1- or 2-element) clusters. However, pilot-study results showed us that it may be more *useful* to have some user-desired number of clusters in order to keep a survey view of a topic, because different information needs may make a coarser or more fine-grained clustering desirable (see Tuzhilin, 2002 for the general necessity of subjective measures of usefulness).

In bibliometrics-based work, the usual procedure for judging cluster validity and usefulness is to ask experts. The questions asked are akin to external and relative measures, but they are necessarily posed qualitatively and elicit answers that involve subjective assessments. Examples are "Do clusters represent specific research topics?" and "Do these topics differ reasonably well among each other?" (Braam et al., 1991), "What was the major impact or implication of [this] article on subsequent research?" and "Could you explain the possible nature of such connections?" (Chen, 2006), the "relatedness" between bibliographically coupled documents (Vladutz & Cook, 1984), or the "relevance" of clusters as new concepts for an ontology (Spiliopoulou, Schaal, Müller, & Brunzel, 2006).

Based on these proposals, we define and obtain the measure *cluster integrity* for a given clustering solution: We first ask experts to label a cluster to describe a research topic. We then ask them to judge, for each element in a cluster, whether it fits that topic or not. The latter serves as a proxy for intra-cluster similarity maximization. Finally, the percentage of fits is averaged over all clusters. In addition, we ask the expert whether and which clusters overlap strongly in content. We divide the number of overlapping clusters by the total number of clusters to obtain the measure of *cluster impurity*.[6]

4.1.1. Method

We asked two domain experts to determine cluster quality.[7] Given the size of the DL and of the set of possible search terms, any choice of search terms for such an evaluation must necessarily be exemplary. We therefore chose 10 search terms that we considered broad and semantically ambiguous enough to produce distinct subtopics (see Table 1).

To find an approximation of a "useful" clustering solution, we considered cognitive capacity: It is well-known that the number of information "chunks" that people can handle simultaneously is limited (see the classic article by Miller (1956) and the literature following it). To use an empirically motivated value, we formed the average of the numbers of document groups that our test users settled on in their final organisation of results (see Section 4.2). The rounded average number of clusters was seven, which is also in line with Miller's (1956) results.

All clusters were formed from a result set large enough to produce at least 30 non-isolated documents ("RFID": 25, the maximum result set in CiteSeer). Total result set sizes ranged from 55 to 108, which were, by the Yahoo! ranking, also the most relevant results. The tool's default settings (including bibliographic coupling) were used.

4.1.2. Results and discussion

Results are shown in Table 1. From the given titles for the different clusters, one can see that topics range from very general collections to specialized topics. The results also illustrate commonalities and differences between raters. First, the quantitative measures, shown at the top of the table, of cluster integrity and impurity were similar, but not identical. An

[6] These measures and procedure pose some challenges. In particular, it could be argued that if the expert names the cluster and evaluates the fit of its elements, the clustering and the labelling are evaluated simultaneously. However, such dependencies between different parts of an expert's evaluation are probably unavoidable, as an investigation of the questions asked for example by (Braam et al., 1991; Chen, 2006) shows.

[7] Two experts were chosen in accordance with the literature as a trade-off between the need to validate results and the high costs of obtaining expert opinions; cf. the two-expert settings used in (Chen, 2006; Lu, Janssen, Milios, Japkowicz, & Zhang, 2007) or even in the highly professionally organised TDT evaluation on topics in news (Cieri et al., 2002). Zeng, He, Chen, Ma, and Ma (2004) used three experts, but they investigated sub-topics of general Web queries, which are easier to judge than scientific sub-topics. Other studies used only one expert and/or unspecified ways of obtaining expert judgments (Braam et al., 1991; Spiliopoulou et al., 2006; Janssens et al., 2008).

Please cite this article in press as: Berendt, B., et al. Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.08.002

Table 1
Experts E1's and E2's assessments of the seven system-generated clusters for different search terms. E2's assessments are given in italics and parentheses. Top: Cluster sizes and quantitative quality measures for 10 search terms. Bottom: Cluster labels for the first 4 search terms.

Search Term	Cluster sizes	Cluster integrity	Cluster impurity
Web mining	6;5;4;5;5;5;5	.80 *(.83)*	.29 *(.57)*
Information retrieval	3;8;4;5;5;3;4	.50 *(.57)*	0 *(0)*
RFID	2;3;3;4;5;4;5	.81 *(.86)*	.29 *(.43)*
Semantic Web	4;6;5;5;5;5;5	.61 *(.73)*	.29 *(0)*
Cluster	6;4;5;3;5;5;4	.46 *(.62)*	.29 *(.29)*
Data mining	4;3;7;5;5;6;6	.59 *(.62)*	.29 *(0)*
Grammar	3;2;8;3;5;5;4	.81 *(.79)*	0 *(0)*
Kernel	5;7;6;5;5;7;6	.83 *(.85)*	.29 *(.43)*
Machine learning	5;5;6;5;5;7;6	.57 *(.56)*	0 *(.29)*
Network	4;5;5;4;5;6;5	.60 *(.61)*	.29 *(.29)*

Web mining
Personalization by usage mining; Web log mining, pattern discovery; modelling user behaviour by Web usage mining; Tools und data preparation; Semantic Web; usage patterns, structure mining; structure analysis *(Personalization, clustering, usage mining; Web usage mining, logs; clustering, user behaviour mining; pattern discovery; Semantic WM; navigation patterns,Web log mining; Web usage mining).*

Information retrieval
User interface; text & linguistic processing; distributed scalable architecture; basics, Web IR; private IR; text classification; NLP *(User interfaces; linguistics, information extraction; distributed IR, IR models; Web IR; Information theory; multimedia IR; NLP).*

RFID
Cryptography; RFID in museum applications; exploring and mapping; crypto-graphy; cryptography, authentication; object identification; mobile usage *(Cryptoanalysis; RFID applications; mapping and localization; security and privacy; privacy; ubiquitous computing; RFID and the WWW).*

Semantic Web
SW services; search services; applying semantic services, DAML; tags, generating metadata; Web services; usage mining; portal server, migrating to SW *(Web services; views, RVL; SW applications; SW tools; Web publishing, portals; SW mining; ontologies).*

analogous observation can be made about the qualitative cluster labels shown at the bottom of the table (for reasons of space, they are only listed for the first four search terms). The labels show that cluster content was generally perceived in the same way; but that different raters often focused on different aspects (e.g., application area vs. method in the first cluster of the first search term). The existence of such differences makes it difficult if not impossible to establish a "gold standard", and it points to the paramount importance of treating the machine-generated clusters as a startpoint for users' individual and interactive (re-)grouping of documents.

4.1.3. Limitations and future work

The obtained clustering results are useful, but not perfect; clusters arose whose elements could belong to other clusters too, and some topics were broken into several clusters. One possible reason for these suboptimal results is the sparsity of the citation matrix. [8] The situation may be improved by integrating further metadata: the cited documents that CiteSeer catalogues as "not in database", and sources such as DBLP, ACM and Google Scholar. In future work, we also intend to study systematic variations of the parameters (similarity category, similarity measure, clustering procedure, clustering criterion) that might deal with this type of data better.

Larger sets of raters and clusters will be necessary to further investigate cluster quality and usefulness and inter-rater agreement, as well as possible specific fits between user groups and methods (Spiliopoulou et al., 2006) or fields and methods.

4.2. Usability and cognitive support

4.2.1. Method

15 graduate students with some experience in online literature search worked with the Web-based version of the tool and answered questionnaires. [9] The questionnaire contained 21 statements to be assessed on a five-point Likert Scale (ranging from "Strongly agree" to "Strongly disagree"), measuring the standard dimensions of usability (Lewis, 1995; Lund, 2001): efficiency, ease of learning, control, usefulness, and satisfaction. In addition, it contained 19 questions on which functionalities were and which additional ones would be considered most helpful.

We let one third of the participants (for technical reasons: 4) use a reduced tool version ("control condition"). This had the same interface as the full version, but did not allow participants to cluster, group, or obtain keywords. The remaining 11 used the full tool with its default settings ("experimental condition"). (This design reflects the observation by Nielsen & Molich (1990) that 3–5 users generally suffice for a heuristic and formative evaluation.)

[8] Our local database contains 716,772 documents (out of CiteSeer's 767,558, which have been fixed as of 2008) and an average of only 2.44 citations per document. Similar sparsity can be observed in other citation datasets such as the INEX 2003 collection, see http://inex.is.informatik.uni-duisburg.de:2003. More recent documents and/or citation extraction algorithms appear to reduce sparsity slightly (April 2009 figure for CiteSeerX: 19.23).

[9] Materials and details are available at http://www.cs.kuleuven.be/~berendt/Bibliography/.

Please cite this article in press as: Berendt, B., et al. Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.08.002

Table 2
Indicators of grouping and mind-map use, quality, and relatedness in tasks 1 and 3: total numbers (averages and, in parentheses, standard deviations) and proportions.

Measure	Task 1	Task 2
# Groups	5.00 (3.38)	6.30 (2.24)
Hierarchical structure:		
# Concepts/ # top-level concepts	2.63 (1.24)	2.47 (1.83)
# Concepts/ # groups	2.90 (1.83)	1.40 (0.84)
Proportion of participants who ...		
Formed groups with high cluster quality	.27	.89
Formed meaningful mind maps	.73	1
Used keywords/labels from the groups	.64	.89
Named All or Some of their clusters	A:.27 + S:.09	A:.22 + S:.11

Students in both groups were first given a task in which grouping was not mentioned and then a task in which grouping was encouraged (literature search for a course essay or publication without/with the instruction to present aspects and subareas of the search term). For about 1.5-2 hours, participants worked on two search terms from the list in Table 1 above. Instructions were given to structure the searches and make them comparable. After the completion of each task, participants were asked to write down a mind map or list (in the following: "mind map") summarizing their results. Participants were then asked to fill out the questionnaire.

Due to the small size of the participants groups, only a descriptive statistical analysis of the quantitative results was conducted. An expert judged the quality of the groupings and mind maps.

4.2.2. Results and discussion: Usability

Participants of the experimental group largely agreed that the tool was usable (median rating of all items of a usability dimension, reversing ratings of negatively-phrased items; averaged over all items of a usability dimension): satisfaction (2.5), usefulness (2.33), ease of learning (2.17), control (3.5), and efficiency (2.67).[10] In particular, most participants found the grouping options "helpful for getting to know new topic areas" (82%) and said they would "prefer this program to [their] previous way of searching" (55%). They also appreciated the other non-standard search functionalities (deleting from the result set, saving results for further processing, ...). Control group opinions indicated that it was not the new tool *per se*, but specifically the grouping functionality that led to the good ratings. For example, no-one preferred the reduced tool over their previous way of searching, and 50% said that grouping would be a helpful new feature.

4.2.3. Results and discussion: Groupings

All participants in the experimental condition used the clustering tool and the opportunities to delete and re-group documents extensively, even in task 1 that contained no specific reason to do so. They often re-clustered several times. All participants in both conditions produced (usually hierarchical) mind maps of the search term topics.

Table 2 shows results on document-group and mindmap-concept numbers and quality. Reported proportions are relative to the number of participants who produced groups and mind maps in the respective task (11 in task 1, 9 in task 2). The results indicate that grouping was used more extensively in task 2 than in task 1 and that the degree of hierarchical structuring increased, both within the mind maps' concepts and between mind-map concepts and cluster-led groups. Also, the quality of both groupings and mind maps increased, and keywords and labels from the cluster-led solution were re-used more extensively for the mind maps in the second task. The re-use of keywords and labels was observed to be meaningful, especially in the second task, and the participants with good groupings also created meaningful mind maps. The changes between the tasks are evidence of learning, including a transition to using and developing the groupings as a first step towards a high-level domain model.

4.2.4. Limitations and future work

These results cannot establish whether people obtained better conceptual structures of the domain of their search term than they would have done without the automatic grouping. An inspection of the mind maps showed that the used method had left many degrees of freedom and introduced noise. In order to test the strong claim of tool usefulness, one would need to confront all participants with a topic about which they know little, subject them to pre- and post-tests of knowledge, and allow for significantly more time for in-depth topic researching.

5. Conclusions and outlook

In this paper, we have proposed a general system architecture and a concrete tool as part of such a system for supporting scientific authors in their use of, and contribution to, Web-based science. The tool focuses on the "reading" phases of authoring

[10] The latter problems were alleviated by improvements to the interface and by changing to a different output technology.

Please cite this article in press as: Berendt, B., et al. Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.08.002

(search/retrieval and sense-making), encouraging authors to actively construct and re-construct literature lists and domain models, and to engage in discussion. Using Yahoo! and CiteSeer, the tool offers a grouping of literature using bibliographic-coupling, co-citation and textual similarity which can be changed, tagged and re-used by the tool's users.

Evaluation studies showed that the interactive and constructive nature was welcomed and seen as a chance to learn more about metadata, citations, and the "web of science". We argued that the judgment of clusters and document groups constructed from them must involve subjective criteria, and showed that clusters and groups represent identifiable sub-topics.

In future work, we plan to professionalize the system and develop a workflow for keeping the system and its use of other resources up to date. For example, the tool is being updated to work with CiteSeerX (http://citeseerx.ist.psu.edu), which went online after the bulk of the work described here was done. CiteSeerX is currently (April 2009) in beta stage, and the major search engines index some combinations of CiteSeer and CiteSeerX. The functionalities relevant for our system have not changed, so a migration is rather straightforward.

In addition, we aim to extend functionality, in particular discussion. Currently, we only encourage the user to assign *some* label to a group of literature. This labelling is a form of "Web2.0" tagging. We plan to combine such tagging and a more traditional form that could be termed "referential tagging": the texts around citations or on citations (anchor texts), cf. Bradshaw (2003). The combination of personal sense-making, referential tagging, and Web2.0 tagging promises to lead to the next generation of intelligent authoring tools.

Acknowledgements

We thank Lee Giles and Isaac Councill for providing us with the CiteSeer code and many answers to our questions.

References

Bani-Ahmad, S., Cakmak, A., Özsoyoglu, G., & Al-Hamdani, A. (2005). Evaluating publication similarity measures. *IEEE Data Engineering Bulletin, 28*(4), 21–28.

Berendt, B., Dingel, K., Hanser, C. (2006). Intelligent bibliography creation and markup for authors: A step towards interoperable digital libraries. In *Proc. ECDL. Vol. 4172: LNCS* (pp. 495–499). Springer.

Bichteler, J., & Eaton, E. (1980). The combined use of bibliographic coupling and cocitation for document retrieval. *JASIST, 31*(4), 278–282.

Bier, E. A., Good, L., Popat, K., Newberger, A. (2004). A document corpus browser for in-depth reading. In *Proc. JCDL* (pp. 87–96). ACM.

Braam, R., Moed, H., & van Raan, A. (1991). Mapping of science by combined co-citation and word analysis I. *JASIS, 42*(4), 233–251.

Bradshaw, S. (2003). Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *Proc. ECDL 2003. Vol. 2769: LNCS* (pp. 499–510). Springer.

Chen, C. (1999). *Information Visualization*. London: Springer.

Chen, C. (2003). *Mapping scientific frontiers: The quest for knowledge visualization*. London: Springer.

Chen, C. (2006). Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST, 57*(3), 359–377.

Chen, C., Carr, L. (1999). Visualizing the evolution of a subject domain: A case study. In IEEE visualization (pp. 449–452).

Cieri, C., Stressel, S., Graff, D., Marey, N., Rennert, K., & Libermann, M. (2002). Corpora for topic detection and tracking. In J. F. Allan (Ed.), *Topic detection and tracking* (pp. 33–66). Berlin: Springer.

Cutting, D. R., Pedersen, J. O., Karger, D. R., Tukey, J. W. (1992). Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proc. SIGIR* (pp. 318–329). ACM.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.) (1996). Advances in knowledge discovery and data mining. AAAI/MIT Press.

Feng, L., Jeusfeld, M. A., & Hoppenbrouwers, J. (2005). Beyond information searching and browsing: Acquiring knowledge from digital libraries. *Information Processing & Management, 41*(1), 97–120.

Fortuna, B., Grobelnik, M., & Mladenic, D. (2005). Visualization of text document corpus. *Informatica (Slovenia), 29*(4), 497–504.

Fortuna, B., Mladenic, D., Grobelnik, M. (2006). Semi-automatic construction of topic ontologies. In M. Ackermann et al. (Ed.), *Semantics, Web and mining. EWMF/KDO workshops at ECML/PKDD 2005. Vol. 4289: LNCS* (pp. 121–131). Springer.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002a). Cluster validity methods: Part I. *SIGMOD Record, 31*(2), 40–45.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002b). Clustering validity checking methods: Part II. *SIGMOD Record, 31*(3), 19–27.

Haveliwala, T. H., Gionis, A., Klein, D., Indyk, P. (2002). Evaluating strategies for similarity search on the Web. In *Proc. WWW* (pp. 432–442).

Janssens, F., Glänzel, W., & De Moor, B. (2008). A hybrid mapping of information science. *Scientometrics, 75*(3), 607–631.

Janssens, F., Leta, J., Glänzel, W., & De Moor, B. (2006). Towards mapping library and information science. *Information Processing & Management, 42*, 1614–1642.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*, 10–25.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction, 7*(1), 57–78.

Lu, W., Janssen, J. C. M., Milios, E. E., Japkowicz, N., & Zhang, Y. (2007). Node similarity in the citation graph. *Knowledge Information System, 11*(1), 105–129.

Lund, A. M. (2001). Usability interface – measuring usability with the USE questionnaire. Retrieved August 25, 2006, from URL http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.html.

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, et al. (2002). On the recommending of citations for research papers. In *Proc. CSCW* (pp. 116–125). New York, NY: ACM.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Nielsen, J. Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proc. CHI-1990* (pp. 249–256). New York, NY, USA: ACM Press.

Qu, Y., & Furnas, G. W. (2008). Model-driven formative evaluation of exploratory search: A study under a sensemaking framework. *Information Processing & Management, 44*(2), 534–555.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *JASIS, 24*(4), 265–270.

Small, H. (1994). A SCI-MAP case study: Building a map of AIDS research. *Scientometrics, 30*, 229–241.

Small, H., & Greenlee, E. (1980). Citation context analysis of a co-citation cluster: Recombinant-DNA. *Scientometrics, 2*(4), 277–301.

Small, H., & Griffith, B. (1974). The structure of scientific literatures, I: Identifying and graphing specialities. *Science Studies, 4*(1), 17–40.

Spiliopoulou, M., Schaal, M., Müller, R. M., Brunzel, M. (2006). Evaluation of ontology enhancement tools. In M. Ackermann et al. (Ed.), *Semantics, Web and mining. EWMF/KDO workshops at ECML/PKDD 2005. Vol. 4289: LNCS* (pp. 132–146). Springer.

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston/ MA: Addison-Wesley.

Tho, Q. T., Hui, S. C., & Fong, A. C. M. (2007). A citation-based document retrieval system for finding research expertise. *Information Processing & Management, 43*(1), 248–264.

Please cite this article in press as: Berendt, B., et al. Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.08.002

Tuzhilin, A. (2002). Usefulness, novelty, and integration of interestingness measures. In *Handbook of data mining and knowledge discovery*. Oxford University Press.
Twidale, M. B., Gruzd, A. A., & Nichols, D. M. (2008). Writing in the library: Exploring tighter integration of digital library use with the writing process. *Information Processing & Management, 44*(2), 558–580.
Vladutz, G., & Cook, J. (1984). Bibliographic coupling and subject relatedness. *Proceedings of the American Society for Information Science, 21*, 204–207.
Zeng, H. -J., He, Q. -C., Chen, Z., Ma, W. -Y., Ma, J. (2004). Learning to cluster Web search results. In M. Sanderson, K. Järvelin, J. Allan, P. Bruza (Eds.), *SIGIR* (pp. 210–217). ACM.
Zhang, X., Qu, Y., Giles, C. L., Song, P. (2008). Citesense: supporting sensemaking of research literature. In *Proc. CHI '08* (pp. 677–680). New York, NY: ACM.

Please cite this article in press as: Berendt, B., et al. Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.08.002