Research literature clustering using diffusion maps

Paavo Nieminen, Ilkka Pölönen*, Tuomo Sipola

Department of Mathematical Information Technology, P.O. Box 35 (Agora), FI-40014 University of Jyväskylä, Finland

Abstract

We apply the knowledge discovery process to the mapping of current topics in a particular field of science. We are interested in how articles form clusters and what are the contents of the found clusters. A framework involving web scraping, keyword extraction, dimensionality reduction and clustering using the diffusion map algorithm is presented. We use publicly available information about articles in high-impact journals. The method should be of use to practitioners or scientists who want to overview recent research in a field of science. As a case study, we map the topics in data mining literature in the year 2011.

Keywords: knowledge discovery process, literature mapping, data mining, clustering, diffusion map

1. Introduction

A task that researchers in any field of science face is to gain an understanding of what others are doing on the field and how it is currently developing. This is a necessary step when relating the researcher's own work to the bigger picture. The research presented here originates from our interest to answer the following basic questions:

1. What main topics are discussed in current data mining research literature?
2. What are the most frequently mentioned methods in the literature?
3. Which journals publish the different topics within the field of data mining?

Very soon we found out that data mining is a rapidly expanding branch of science with a large number of articles published about it each year. Therefore, gaining a general view about the publication space turns out to be, in practice, quite challenging.

A rigorous way to create a secondary study would be to perform a systematic literature review. Originating from medical sciences, systematic reviews can be used also in other disciplines, exemplified by the adaptation to software engineering by Kitchenham

*Corresponding author.

Email addresses: paavo.j.nieminen@jyu.fi (Paavo Nieminen), ilkka.polonen@jyu.fi (Ilkka Pölönen), tuomo.sipola@jyu.fi (Tuomo Sipola)

(2004). A systematic literature review creates a synthesis about a specific phenomenon by conglomerating the evidence published in primary research papers. There is also a lighter version of systematic literature review called mapping study, or scoping review, that intends to identify groups of current literature and identify gaps for further, more detailed, literature review (Budgen et al., 2008). Mapping study, even if lighter than a systematic review, is still a laborious task to do for a massive body of literature.

As data mining methodologies facilitate the handling of huge data masses, it would seem natural to use them to summarize the research literature itself. After all, a definition of data mining, according to Hand et al. (2001, p. 1), is "*the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.*" As it turns out, others have followed a similar way of thinking and studied the creation of automated tools for literature surveys. For example, Cohen et al. (2006), and later Matwin et al. (2010), use machine learning algorithms to assess the relevance of articles in order to reduce the workload of experts who maintain systematic reviews common in evidence-based medicine.

Our goal greatly resembles those pursued by researchers in the field of scientometrics, which is commonly defined as the quantitative study of science. Ivancheva (2008) provides a categorization of scientometrics methodology for research subjects, information types and method classes. Our research subject can be seen as *science by itself* because we try to understand the structure of a field of science. The field is limited, focused and concrete, so the information type of this research is *operational*. Finally, in the classification of Glänzel (2003) our work positions itself in *structural scientometrics* trying to map the research area.

Classical methods used in science mapping, for example in planning of research policies or finding out structures in scientific communities, include those of co-citation analysis (Small, 1973) and co-word analysis (Callon et al., 1983). Co-citation analysis looks for structure in research literature by analyzing the frequency that an article is cited together with another one in later works. Co-word analysis is based on the idea that the text in scientific publications connects key concepts to each other. In co-word analysis, connections between the concepts emerge from the network of co-occuring words instead of the network of citations made between authors.

For the goal of mapping literature, metadata could be used instead of the full research papers. Metadata is usually more readily available and, additionally, it should contain less noise because it is very focused in content and limited in form. There are existing metadata and article databases for certain fields of science. Some of the more notable examples are CiteSeerX[1], DBLP[2], arXiv[3] and PubMed[4]. CiteSeerX is an online database that collects article metadata focusing primarily on the computer and information sciences. DBLP is a database for computer science focusing on authors. ArXiv covers mathematics, computer science, nonlinear sciences, quantitative biology

[1] `http://citeseerx.ist.psu.edu/`
[2] `http://www.informatik.uni-trier.de/~ley/db/`
[3] `http://arxiv.org/`
[4] `http://www.ncbi.nlm.nih.gov/pubmed/`

and statistics. PubMed archives biomedical literature citations. There are also existing software frameworks to collect information about scientific articles, for example that of CiteSeerX (Teregowda et al., 2010), using web spider technology and various heuristics to collect metadata and citations. The original article databases, and the metadata repositories, can be accessed via web browser interfaces and in some cases also machine-readable interfaces such as the OAI protocol[5]. A major interdisciplinary database with a significant role in the development of scientometrics is the Thomson Reuters (formerly known as ISI) Web of Knowledge (WoK)[6].

To utilize these databases efficiently, computational methods are required. Current work about literature database analysis seems to focus on analyzing citations. One example of such a system is CiteSpace that finds trends and patterns in scientific literature. It was tested with mass-extinction research and terrorism research (Chen, 2006). There have also been schemes for recommending research papers using citation data with subspace clustering based analysis (Agarwal et al., 2005).

Journal interdisciplinarity has been studied with citation reports by clustering using bi-connected graphs (Leydesdorff, 2004). Leydesdorff & Rafols (2009) used factor analysis to cluster the ISI subject categories. Later, these results were replicated for the revised list of categories (Leydesdorff et al., 2013). The methods can be used to produce global maps of sciences, which are two-dimensional illustrations of global literature, in which subsets such as the publications of researchers or companies can be positioned and compared with each other (Rafols et al., 2010).

Tseng & Tsay (2013) present a data processing pipeline that identifies subfields of science. With Dice coefficient similarity and multi-stage clustering, they cluster journals. They believe that articles form topics or categories which in turn form subfields. The research uses manual cluster labeling, but the task is assisted with text mining. The results include subfield descriptions and visualizations of topical maps.

Crimmins et al. (1999) use their framework to discover information from the Internet. They collect frequently occurring phrases, citation and metainformation, summarizing the results into a contingency table. The framework provides clustering and principal component analysis capabilities. Clustering and visualization produce maps that facilitate the understanding of the searched information. This kind of approach seems reasonable also in the context of scientific articles, because there is a similar graph-like structure.

As further examples, clustering frameworks for more traditional text mining have been used to analyze large text databases. Bravo-Alcobendas & Sorzano (2009) clustered biomedical papers using non-negative matrix factorization and k-means algorithms. Aljaber et al. (2010) used various clustering methods to examine literature concerning high energy physics and genomics. Their datasets are from knowledge discovery competitions and workshop tasks[7]. They show that the combination of citation information and extracted features from full article text produces an efficient way to capture the content of scientific papers.

[5] `http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm`

[6] `http://wokinfo.com/`

[7] KDD Cup 2003, TREC 2006 and 2007 Genomics Tracks

We view computer-assisted literature mapping as a special case of the process of knowledge discovery in databases, as described by Fayyad et al. (1996a,b), and we shall continue using terminology related to their description, which is presented in Figure 1. The steps from raw data to the goal (knowledge to be discovered) involve selection, preprocessing, transformation and mining of the data, as well as representing and interpreting the discovered patterns. The goal in our case is not so much to aid in matters of policy, but to help a researcher gain an initial understanding of what others are currently doing in the same research field. Therefore, we are interested in applying data mining to the concepts (keywords) being discussed in the literature rather than the authors and their affiliations. The electronic articles that reside in databases owned by journal publishers form the bulk of raw data. Consequently, we want keyword vectors to be the transformed data.

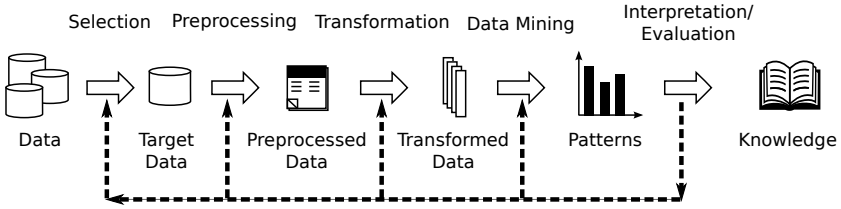

Figure 1: Steps of the knowledge discovery process after Fayyad et al. (1996a).

The technical data mining steps used by Szczuka et al. (2012) in their document grouping and concept identification system are similar to those used in our approach. However, we build upon the clustering approach by using a diffusion map dimensionality reduction step. In addition, our case study analyzes a somewhat larger number of articles. These articles are a subset of scientific literature, and are selected using a specified procedure. Our features are based on the publicly available metadata, while Szczuka et al. use the whole text of the articles, which is feasible when they are easily available.

In this paper, we propose a knowledge discovery and data mining method to create a global view of current topics in a particular field of science using publicly available information about publications in high-impact journals. We compare recent articles using their keywords and title words using a diffusion map data mining approach. The purpose is to find the current snapshot state and structure of the research field based on the data. Maps of science are mostly built upon citation information, but the interests of this article lie in the content of the articles, not connections of citations. In Section 2 we describe the details of our approach, and adapt it to our case study in Section 3. Section 4 presents and discusses the results regarding the data mining literature case study. Section 5 provides a summary of this research.

2. Methodology

We present a clustering framework, which is designed to be useful when searching for a general overview of topics covered in a body of text documents. The major steps

in our metadata-based clustering framework follow the adapted knowledge discovery process (Fayyad et al., 1996a,b). The adapted steps include:

1. Selection of relevant literature.
2. Dataset formation (preprocessing and transformation).
3. Data mining the article set with dimensionality reduction and clustering.
4. Interpretation of the summaries obtained from the previous step.

Later on, in Section 3, we present our procedure using data mining literature as an example. However, the steps are in no way limited to any specific field of research that one might want to study.

2.1. Selection of relevant literature

The first step of the process, i.e., selection of the relevant research literature, is important because it defines the publication space. These steps could be automated but at least some initial query from the user must restrict the search. We suggest the following general steps:

1. Identify journals that are likely to be relevant to the field of interest.
2. Focus on the most relevant journals within the identified ones.
3. Decide on further restrictions, e.g., dates of publication.

How this selection is done depends on the research goals. Subsequently, in Section 3, we make suggestions that are based on our experiences and could be used when the goal is similar.

2.2. Article dataset formation

After selecting the body of literature to be studied, metadata needs to be gathered and preprocessed. The main steps, which should mostly be automated, include the following:

1. Gathering data.
2. Normalization of data.
3. Feature extraction.
4. Construction of feature matrix.

Gathering data may be done in various ways, e.g., web scraping, accessing public databases or using public APIs. Data normalization consists of unifying notational conventions and spelling. Feature extraction gathers numerical features from the available textual information. As a final step, a feature matrix is constructed for data mining. The dimensions of this matrix are $n_{\text{articles}} \times n_{\text{features}}$.

2.3. Data mining

In what follows we describe our data mining and analysis steps consisting of article clustering, keyword frequency counting and computation of journal distribution within clusters.

2.3.1. Article clustering

After preprocessing and matrix formation, the data is clustered in order to look for the most dominating groups of topics. The overall procedure of article clustering consists of two steps:

1. Dimensionality reduction using diffusion map.
2. Clustering using agglomerative method with Ward distance.

The first step produces an eigenvector presentation of the transition matrix of the data. This presentation reduces noise in the data, makes the clustering easier and enables visualization. The second step is a simple clustering task.

The binary matrix obtained from data formation step can be of high dimensionality, for example in the order of thousands. In bibliometrics and scientometrics this problem is commonly solved with a combination of hierarchical clustering and multidimensional scaling (MDS) for dimensionality reduction (Boyack et al., 2005; Waltman et al., 2010). Our approach is fundamentally the same, but instead of MDS we employ the diffusion map algorithm (Coifman & Lafon, 2006). It finds a low-dimensional representation using the singular value decomposition of a transition probability matrix based on some chosen distance function. Thus, the high-dimensional data points become embedded in a lower-dimensional space. The dimensionality reduction yields a space where the Euclidean distance corresponds to the diffusion distance in the original space (Coifman & Lafon, 2006; Nadler et al., 2008).

Let us consider a dataset $X = \{x_1, x_2, \ldots, x_n\}, x_i \in \{0, 1\}^p$, that consists of vectors of binary digits, where n is the number of data points and p is the number of measured features. The initial step of the diffusion map algorithm calculates the affinity kernel matrix W, which has data vector distances as its elements:

$$W_{ij} = \exp\left(-\frac{\mathrm{dist}(x_i, x_j)}{\epsilon}\right),$$

where $\mathrm{dist}(x_i, x_j)$ is the similarity measure of Jaccard (1901). Our algorithm uses this for the initial distance matrix between the articles, because only the non-zero elements should contribute to the distance metric. A kernel is used in order to bring close points closer and to increase the distance to distant points.

The row sum diagonal matrix $D_{ii} = \sum_{j=1}^{n} W_{ij}, i \in 1 \ldots n$ is used to normalize the W matrix: $P = D^{-1}W$. This matrix represents the transition probabilities between the data points. The conjugate matrix $\tilde{P} = D^{\frac{1}{2}} P D^{-\frac{1}{2}}$ is created in order to find the eigenvalues of P. With substitution we get

$$\tilde{P} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}.$$

This normalized graph Laplacian (Chung, 1997) preserves the eigenvalues (Nadler et al., 2008). Singular value decomposition (SVD) $\tilde{P} = U \Lambda U^*$ finds the eigenvalues $\Lambda = \mathrm{diag}([\lambda_1, \lambda_2, \ldots, \lambda_n])$ and eigenvectors $U = [u_1, u_2, \ldots, u_n]$ for $\tilde{P}$. The eigenvalues for P are the same as for $\tilde{P}$. The eigenvectors for P are found with $V = D^{-\frac{1}{2}} U$ (Nadler et al., 2008). The low-dimensional coordinates Ψ are created using $\Psi = V\Lambda$. Only a few of these coordinates are needed to represent the data to a certain degree of error (Coifman & Lafon, 2006).

Basically, the row-stochastic Markov matrix P corresponds to modes of a random walk on the data. It should be noted that the eigen-analysis is based on the distance matrix rather than the data matrix. The use of the kernel brings the neighborhood closer to the point. Points that are close to each other on the graph are also close in the embedded space. Diffusion map has a fundamental difference to principal component analysis (PCA) and multi-dimensional scaling (MDS) there: it also reveals nonlinear relationships between features in embedded space. Linear projections (PCA and MDS) cannot show these.

Diffusion map facilitates the clustering by simplifying the representation of data. Therefore, simple clustering methods can be used to find relevant structure of the data. For clustering the articles using the low-dimensional coordinates, we apply agglomerative clustering using the Ward method for cluster distances. The agglomerative hierarchical clustering scheme is discussed in Everitt et al. (2001, ch. 4) and Hastie et al. (2011, p. 523). The number of clusters is determined using the silhouette measure; the number yielding the highest average silhouette for a clustering is chosen, as recommended by Rousseeuw (1987). When compared to the brief overview by Waltman et al. (2010) our combination of diffusion map dimensional reduction and clustering seems to be unique in the field of science mapping, although it is previously shown to be both theoretically sound and applicable to many real-world tasks, including document clustering (Lafon & Lee, 2006).

The clustering usually has a dense center forming one cluster and a few sparser clusters that stand out. For this reason, the clustering was repeated using only the remaining center, which we call the residual cluster. We end up with an overall iterative data clustering method that includes the following steps:

1. Dimensionality reduction using diffusion maps.
2. Agglomerative clustering with optimal silhouette.
3. Take small clusters as results, and remove them from further analysis.
4. Continue from step 1 using the big residual cluster until stopping criterion is met.

2.3.2. Keyword frequency and journal distributions

Simple keyword analysis helps to identify the topics that have been discussed the most in the examined set of articles. The number of how many articles include each keyword is counted. A simple sum over all the articles yields overall keyword frequencies. In our case study, the purpose of this step was to find out the most common methods and applications in current literature.

As yet another additional piece of information, we compute the distribution of journals in the clusters. Each article in a cluster belongs to a single journal and it is easy to create a frequency table. This table supports the knowledge discovery task by showing the relations between the generated clusters and the journals.

2.4. Interpretation

The data mining analysis step produces summaries of the data which need to be interpreted by the user. They can be presented in the form of visualizations, tables and lists. The evaluation of the results depends on the initial search goals. It is up to

the user to decide whether the obtained clustering, structural visualization and found categories are sensible. We do this verification by comparing the results with published expert opinions.

2.5. Comparison with other scientometric methods

A short comparison with other analysis methods is provided, because the reader might not be familiar with our approach.

Traditional co-word analysis compares word pairs found in literature. The pairs are created from the body of literature and the co-occurrence frequencies are collected to a matrix (Callon et al., 1983). These words and their relations are believed to define concepts in the scientific field. The concepts can be connected and clustered using graph algorithms. However, the approach described in our research clusters *articles*, not word pair concepts. We measure the distances between articles using keywords. Naturally, word co-occurrence plays a part also in our method via the chosen Jaccard distance metric and the diffusion process modelled by the dimension reduction algorithm.

OpenOrd is a highly scalable citation graph based method for science mapping (formerly known as as VxOrd or DrL), used by Boyack et al. (2005). OpenOrd uses state of the art graph algorithms to produce (x, y) -coordinates and pruned edge distances for the articles being examined. Standard clustering methods, such as k-means can then be used to find structure in the data. Albeit similar, our method differs in three major ways. First, we use keywords instead of citations in the similarity matrix computation. Second, the optimization problem being solved is different. Visualization methods try to optimize for clarity, while diffusion map aims to retain the diffusion distance. Third, the dimensionality of our output space can be more than two, as our main goal is clustering rather than distance visualization.

3. Adaptation for the case study

This section presents an adaptation of the methodology for the case study. The abstract steps introduced in Section 2 are now applied to current data mining literature. Figure 2 shows the adapted knowledge discovery steps to fit the task of mining specified literature. Each step now contains more phases and the detailed execution has to be determined. The redefined steps are as follows:

1. Selection of relevant literature using impact factors and manual screening of journals.
2. Dataset formation (automatic preprocessing and transformation), including web scraping, filtering, normalization and title conversion.
3. Data mining with dimensionality reduction and clustering.
4. Interpretation of the summaries obtained from the previous step and comparison with published expert opinions.

These steps are detailed in the following subsections and the motivation behind them is discussed.

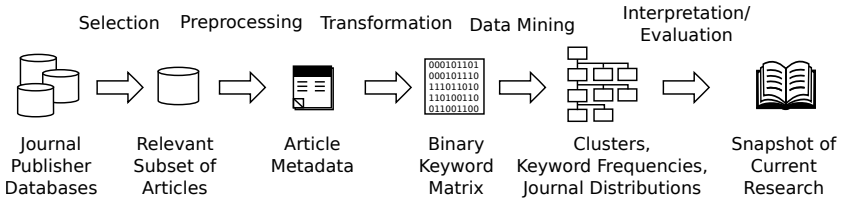

Figure 2: Our adaptation of the knowledge discovery process for mapping research literature based on Fayyad et al. (1996a), cf. Figure 1.

3.1. Selection of relevant literature

In this practical case the literature selection step in the methodology is specialized to find the most relevant journals and articles. In order to limit our data to only articles concerning the field of data mining, we used the following restrictions:

1. Selecting journals that are listed in WoK.
2. Limiting the WoK subject categories.
3. WoK impact factor over a threshold.
4. Further voting about the relevance to data mining.
5. Limiting the target time frame.

To identify relevant journals, we suggest using the impact factor metric published yearly in the Journal Citation Report[8] of WoK. Impact factor (Garfield, 1972) is a numerical value, that provides a quantitative tool for ranking journals based on their impact to a field of science. The impact factor is computed by dividing the number of citations made to the articles of a journal by the total number of articles published during a time window. Longer-term impact factors and trend graphs are available from WoK, but we restricted our scope to one-year impact factors in order to get a recent picture of the quickly developing field of data mining, with the newest journals included. Impact factors of 2010 were the most recent ones available when starting our work.

Despite its limitations and pitfalls, discussed, for example, in Seglen (1997), impact factor is regarded as a *de facto* tool for assessing the relevance of journals. Therefore, we chose to restrict our study to journals with impact factor higher than the arbitrarily selected threshold of 1.0 in order to focus only on the most cited research. Comparing impact factors might not generalize to very interdisciplinary topics, because the metric is not comparable across fields of science due to different citation cultures. However, in our case of data mining, we expect the topic to be covered mostly by journals focusing on computer science, statistics and mathematics, between which we expect the citation culture to be similar.

The Thomson Reuters Web of Knowledge divides the listed journals to 176 subject categories. Not all of these categories are related to the field of science that is in focus. In our study we selected the following categories, which in our opinion should contain

[8] `http://wokinfo.com/products_tools/analytical/jcr/`

most of the work related to the field of data mining: “computer science, artificial intelligence”; “computer science, information systems”; “computer science, interdisciplinary applications”; “mathematics, applied”; “mathematics, interdisciplinary applications”; “statistics & probability”.

In Tables 1 and 2, we list the journals that were initially identified as the candidate data sources for this study, i.e., they were listed in the WoK, had an impact factor of at least 1.0 and included one of the words *Data Mining (dm)*, *Data Engineering (de)*, *Knowledge Discovery (kd)*, *Knowledge Engineering (ke)* or *Data Analysis (da)* in their editorial statements or public scope definitions. The technical filtering did not seem to single out the most data mining related journals, perhaps due to the term data mining being a buzzword used more than it factually should. So finally, to focus on the most relevant ones, we voted for inclusion of journals based on inspection of the journals' editorial statements and preliminary browsing of their content. The threshold of inclusion was that at least two of the three authors regarded the journal relevant. In Table 1, we show the journals that were finally selected for inclusion in this study, based on subjective evaluation of each journal's relevance to our research questions. In Table 2, we list the journals that were initially identified but finally rejected. The last column of the tables shows the number of relevance votes that each journal received from the authors.

In our case study, the purpose was to get a snapshot of recent publications, so we chose to restrict our study to the articles published during the year 2011.

3.2. Article dataset formation

We built our database using web scraping to collect data directly from the journal databases via the public WWW interfaces provided by the publishers. Other sources could be added for further studies. For this study the scraper reads the WWW pages of the journal publishers and yields a database entry for each article, including the title, keywords and name of the journal where the article has been published. All published titles from each journal will be listed at this stage, including many non-essential ones, such as editorial comments, letters to the editor, book and software reviews and calls for papers. These non-essential titles are then automatically filtered out based on words contained in the title. Our approach does not extract keywords from the text. Instead, it uses the available metadata and assumes that they are correctly entered by the authors.

Also, some further pre-processing was found to be necessary because of varying formats and conventions found in the data sources. Notational conventions were occasionally found to differ also between different articles within a journal. These discrepancies necessitate a technical cleaning step, where HTML tags are removed, Greek letters and mathematical symbols are converted to corresponding LaTeX expressions, and the separating characters in keyword lists are heuristically chosen. In order to further normalize the keyword lists, we created an automatic tool that converts plurals to singular form, and British English spellings into their American English equivalents.

Feature extraction from the metadata is straightforward. The occurrence of keywords describes the contents of an article, which means that a binary feature vector can be used to represent an article. While inspecting the author-defined lists of keywords,

Table 1: Selected journals after relevance vote.

Selected journal	Scope	Publisher	rel.
ACM Transactions on Information Systems	dm,kd	ACM	2
Applied Soft Computing	dm	Elsevier	2
Bayesian Analysis	dm	ISBA	3
Computational Statistics & Data Analysis	dm,da	Elsevier	3
Computer Journal	dm	Oxf.UP	3
Data Mining and Knowledge Discovery	dm,kd,da	Springer	3
Fuzzy Sets and Systems	da	Elsevier	2
Genetic Programming and Evolvable Machines	dm	Springer	2
IEEE Transactions on Knowledge and Data Engineering	de	IEEE	2
Information Sciences	de,ke	Elsevier	3
International Journal of Approximate Reasoning	da	Elsevier	2
International Journal of Information Technology & Decision Making	dm	World Sc.	2
International Journal of Innovative Computing Information and Control	dm,kd, da	ICIC	2
Journal of Computational and Graphical Statistics	da	ASA	2
Knowledge and Information Systems	dm,de, kd,ke	Springer	2
The Knowledge Engineering Review	ke	Cambr.UP	2
Machine Learning	dm	Springer	3
Pattern Analysis and Applications	ke	Springer	2
Pattern Recognition Letters	dm	Elsevier	3
Statistics and Computing	dm,da	Springer	3

Table 2: Excluded journals after relevance vote.

Excluded journal	Scope	Publisher	rel.
ACM Transactions on Database Systems	dm	ACM	0
ACM Transactions on Internet Technology	dm,kd	ACM	0
ACM Transactions on the Web	dm	ACM	0
Artificial Intelligence in Medicine	ke	Elsevier	0
Computer-aided Civil and Infrastructure Engineering	de	Wiley	0
Computers in Industry	ke	Elsevier	0
Data & Knowledge Engineering	de,ke	Elsevier	0
Electronic Commerce Research and Applications	dm	Elsevier	0
Environmental Modelling & Software	dm	Elsevier	0
Expert Systems with Applications	kd	Elsevier	1
Information Systems	dm	Elsevier	1
Integrated Computer-Aided Engineering	kd	IOS Press	0
Journal of Database Management	dm,ke	IGI Publ.	1
Journal of Hydroinformatics	ke	IWA Publ.	0
Journal of Molecular Modeling	da	Springer	0
Journal of Quality Technology	ke	ASQ	0
Journal of Web Semantics	kd	Elsevier	0
Psychometrika	da	Springer	0
SAR and QSAR in Environmental Research	da	Taylor&Fr.	0
Stata Journal	da	StataCorp	1
World Wide Web – Internet and Web Information Systems	dm	Springer	0

we found out that the keywords, even after normalization, were quite different from each other, even when the articles could have been related to similar topics based on their titles. To improve the situation, our system augments the list of keywords in the following way:

1. List all of the original keywords (for example "face recognition").
2. Add to the list also split, i.e., single-word, versions of the original keywords (for example "face" and "recognition").
3. Remove common English stopwords (such as "a", "the", "in", "and", ...) from the list.
4. Remove also some additional words very common in scientific titles (such as "using", "based", "novel", "new", ...).

Each article is then judged by the software to be related to a keyword in the list if the keyword is found within the title or within one of the keywords of the specific article. For example, an article with the title "About face recognition" would be related to the keywords "face recognition", "face" and "recognition". The information is stored as a binary matrix where each row corresponds to an article and each column to a keyword in our augmented keyword list. A non-zero element means that the keyword is found from the title or keyword list of the article.

At the end of this step, we automatically remove singleton keywords and articles, i.e., keywords that appear only once and articles that contain no keywords common with any other article. Such singleton words are irrelevant in analyzing connections between the articles. In our case study, the final keyword list contained 11,844 words or phrases, and with 2,511 articles the size of the matrix was $2,511 \times 11,844$. After removal of singleton words and articles, 4,187 keywords and 2,499 articles remained. The data matrix of size $2,499 \times 4,187$ was sparse; only 0.3% of its values were ones instead of zeros.

3.3. Data mining

The data mining step follows closely the article clustering approach presented in Section 2.3.1. Figure 3 shows the clustering results for our case study at the first iteration level. The visualization uses the first three dimensions, although empirically chosen first six dimensions were used in the analysis. These coordinates in the figure correspond to the three largest eigenvalues obtained from the diffusion map algorithm.

The iterative approach clusters the articles into several categories, which can be used to analyze the structure of the dataset. The obtained clusters vary considerably in size. Inspection of the keywords and titles in the clusters reveal that the separated clusters have high semantic cohesion. The iteration is stopped when the total size of the separated clusters becomes smaller than 2% of the original data. We conjecture that the most important clusters according to the keyword vectors are found during the first few iterations.

4. Results of the case study

This section presents the results of our case study with key findings and answers to the original research questions: the main topics, most frequent methods and journal

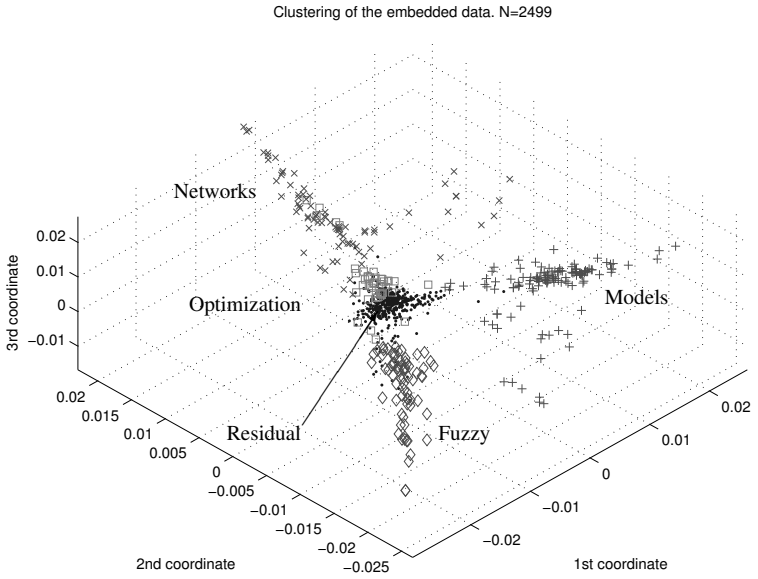

Figure 3: Low-dimensional embedding of the article dataset. Each point corresponds to one article. Different clusters are marked differently, and given their interpreted names. The dense residual cluster can be seen in the middle. For visual clarity, only one third of the data (randomly selected) is plotted in this figure.

identities in the field of data mining, taking into account the restrictions set by the article selection process. We find that the most convenient order is to report the findings from simple keyword frequency counts first, and then to continue with the results from clustering and journal distributions.

4.1. Keyword frequency analysis

Of the 4187 keywords only some were obviously related to data mining methods. This led to a subjective screening of the keywords. The most common method-related keywords and their frequencies were *fuzzy* (327), *optimization* (198), *classification* (172), *clustering* (119) and *Bayesian* (112). All of these are rather general method families.

The more specific method families were not mentioned as frequently. The following list includes notable examples of these method-related keywords: *neural network* (63), *genetic algorithm* (62), *stochastic* (53), *particle swarm optimization* (42) *support vector machine* (40), *fuzzy logic* (36), *feature extraction* (30), *feature selection* (30), *pattern recognition* (27), *evolutionary algorithm* (26), *self-organizing* (23), *decision tree* (19), *genetic programming* (18), *reinforcement learning* (17), *hidden Markov model* (17), *PCA* (16), *differential evolution* (15), *self-organizing map* (14), *dimensional reduction* (14), *least squares* (13), *kernel method* (13), *Kalman* (13), *fuzzy clustering* (12), *k-means* (11), *manifold learning* (11), *feature detection* (8), *c-means* (8) and *independent component analysis* (6).

Some other findings, that were omitted from the above list, are worthy of a short discussion. There were 63 articles that had *data mining* itself as a keyword. The frequencies of keywords *linear* (125) and *non-linear* (61) tell something about the expected result that linear methods are studied or used more widely. Four often mentioned application areas were *face recognition* (32), *wireless sensor network* (30), *image segmentation* (23) and *text analysis* (12).

4.2. Structural view using clustering

The iterative clustering resulted in 19 clusters on five iteration levels and a final residual cluster of size 598. Therefore, 76% of the data falls within these 19 identified clusters. Figure 4 illustrates the levels of the iterative clustering process. The clusters are manually labeled from the most common keywords inside them. On the highest level, the original data of 2,499 articles was clustered into four smaller clusters and a residual cluster of 1,459 articles. We chose a descriptive name for each cluster by examining the 10 most common keywords in the cluster.

Thus, the highest level revealed the following clusters (number of articles in parentheses): *Models* (388), *Networks* (241), *Fuzzy* (239) and *Optimization* (172). The *Models* cluster included also keywords such as *Bayesian*, *fuzzy*, *Markov* and *regression*. The *Networks* cluster covered both *neural networks* and *sensor networks*. The *Fuzzy* cluster included topics such as *fuzzy sets* and *fuzzy logic*. Finally, the *Optimization* cluster included *particle swarm optimization* and topics related to *evolutionary* and *genetic algorithms*.

The second level was obtained by clustering the residual cluster (1,459 articles) of the first level. Clusters on this level were named *Images*, *Learning*, *Face/Pattern*

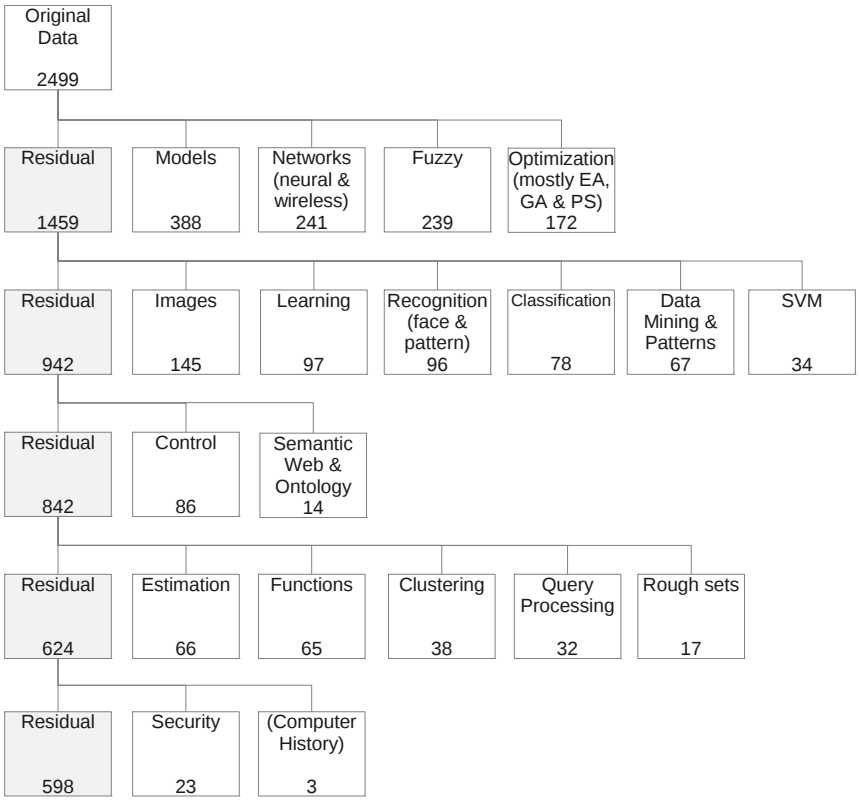

Figure 4: Clusters found during the first five iterations of the algorithm. The numbers tell how many articles fall into the respective clusters. Names are given by inspection of cluster contents.

recognition, *Classification*, *Data mining & Patterns*, and *SVM*. Like on the first level, the descriptive names were chosen on the basis of the 10 most common keywords in the clusters. For example, common keywords in the *Images* cluster contained *image segmentation*, *image retrieval* and *classification*.

The third level extracted two new clusters that we call *Control* and *Semantic web & Ontology*. The fourth level revealed the clusters of *Estimation*, *Functions*, *Clustering*, *Query Processing* and *Rough Sets*. The fifth level yielded one more larger cluster, *Security*, and a very small cluster *Computer History*. The ending criterion was met on this level.

4.3. Journal distribution

The number of articles is not uniformly distributed among the journals, as shown in Table 3. It is also seen that each journal has its own areas of interest with respect to the clusters identified by this study. For example, Pattern Recognition Letters publishes articles related to the clusters *Recognition* and *Images*; in contrast, articles published in Fuzzy Sets and Systems belong to the *Fuzzy* cluster. On the other hand, journals like International Journal of Innovative Computing, Information and Control (IJICIC) and Information Sciences relate to almost all the clusters in the taxonomy discovered by our framework.

4.4. Discussion

This case study presented one viewpoint to understand recent data mining literature. This discussion compares our results to the expert opinion. The advancement of the field has been of interest to the community, and accordingly some overviews have been made. Among recent overview literature there are some interesting papers, such as the one by Kriegel et al. (2007) where the authors envision the major challenges in data mining and knowledge discovery today and especially in the future. Venkatadri & Reddy (2011) give a general overview of current and future trends in data mining. In a similar manner, Kumar & Bhardwaj (2011) review potential future application areas. Wu et al. (2008) give a list of top data mining algorithms based on the opinions of an expert panel. We contribute to this discussion by the quantitative results presented above. Although interesting and enlightening reading, the current reviews and position papers seem to be somewhat restricted in their scope of selected literature, whereas our study attempts to sample the current state of the leading data mining research holistically with an objective, structured and more unbiased method that is based on a methodically selected subset of literature.

The definition of current data mining research is, to an extent, a question of opinion. However, our results seem to adhere to the opinions of other data mining experts. The findings in Section 4.1 about methods are quite similar to KDnuggets poll answers[9], where "academic" persons' most used algorithms in data mining in 2011 were genetic algorithms, support vector machines and association rules. In their brief review, Venkatadri & Reddy (2011) recognize neural networks, fuzzy logic and genetic

[9] `http://www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html`

Table 3: Journal distribution in clusters

	ACM Trans. Inf. Syst.	Appl. Soft. Comput.	Bayesian Anal.	Comput. Stat. Data An.	Data Min. Knowl. Disc.	Fuzzy Set Syst.	Genet. Program. Evol. M.	Inform. Sciences	Int. J. Approx. Reason.	IJITDM	IJICIC	J. Comput. Graph. Stat.	IEEE T. Knowl. Data En.	Knowl. Inf. Syst.	Mach. Learn.	Pattern Anal. Appl.	Pattern Recogn. Lett.	Stat. Comput.	Comput. J.	Knowl. Eng. Rev.
Models	4	15	27	92	6	10	1	35	12	8	51	14	11	4	16	7	33	27	13	2
Networks	1	31	1	2	3	4	2	30	7	1	67	0	14	7	5	1	4	2	59	0
Fuzzy	0	34	0	2	0	64	0	57	18	9	45	0	4	2	0	0	4	0	0	0
Optimization	0	42	1	4	0	0	12	38	0	2	43	0	8	2	3	2	3	1	9	2
Images	1	5	0	5	1	0	1	15	0	1	36	0	2	0	0	8	63	0	7	0
Learning	1	6	0	0	3	0	1	17	2	2	7	1	8	6	21	1	16	0	3	2
Recognition	0	5	0	0	1	0	0	2	1	1	20	0	3	1	0	1	58	0	3	0
Classification	1	4	0	3	3	1	0	8	5	2	11	0	8	5	1	3	20	2	0	1
Data Mining & Patterns	1	5	0	0	6	0	0	14	1	2	6	0	14	15	3	0	0	0	0	0
SVM	0	4	1	3	0	0	0	7	0	0	5	1	1	1	1	1	9	0	0	0
Control	0	3	0	8	0	0	0	7	0	0	63	0	0	0	0	0	2	0	3	0
Semantic Web & Ontology	0	0	0	0	0	0	0	1	1	1	1	0	1	2	0	0	0	0	0	7
Estimation	0	0	0	21	0	1	0	7	2	1	10	3	1	1	0	1	13	5	0	0
Functions	0	1	0	10	0	9	0	15	14	2	8	0	2	2	0	0	2	0	0	0
Clustering	0	0	0	2	0	0	0	6	0	0	4	2	3	3	0	0	16	1	1	0
Query Processing	0	0	0	0	0	0	0	2	0	0	0	0	27	1	0	0	0	0	2	0
Rough sets	0	0	0	0	0	3	0	8	3	0	2	0	0	0	0	0	1	0	0	0
Security	0	0	0	0	0	0	0	8	0	0	12	0	0	0	0	0	0	0	3	0
Computer History	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
Residual	7	8	2	75	8	22	7	71	28	18	116	33	21	23	9	8	57	13	56	16
TOTAL	16	163	32	227	31	114	24	348	94	50	507	54	128	75	59	33	301	51	162	30

programming as the future trends of data mining. Our results in Section 4.2 agree with their findings since corresponding clusters were found already on the first level of our iterative algorithm. Journal distribution analysis in Section 4.3 showed that most journals specialize in just a few topics. However, some journals publishing more diverse topics were also found. The journals adhere to the obtained clustering quite closely, which can help a researcher to select a publication venue.

Overall, our findings seem to agree with the definition of data mining by Hand et al. (2001), which suggests that what is done currently under the label of data mining still studies the problems stemming from the definition given over ten years ago.

4.5. Benefits and limitations

In our study, we did not use an existing benchmark corpus because one main goal of the research was to apply the method to immediately gain new information about recent data mining literature. The method is verified by comparing it to existing expert opinion instead. We wanted to base our study on freely available public data, which excludes full texts in many cases. This unfortunate fact was noted also by some of the researchers we have cited above. The use of full texts would have given a larger feature space and produced more noise. While the main connections might be the same as when using metadata, the additional data mass could have created unforeseen connections between articles that cannot be produced with mere metadata.

To our knowledge this is a unique study of this kind performed on recent data mining literature, which should make the results useful for the data mining community.

5. Conclusion

Following the knowledge discovery process, we created a literature mapping framework based on article clustering. It can be used to analyze topics of current interest in a particular field of science. As a case study, we tested the framework with data mining research literature. Our approach uses publicly available metadata about articles published in high-impact journals. The proposed methodology can be automated, but a more delicate screening may use manual approach in needed steps. In the case study, the data source selection and interpretation included manual work. The methodology is mainly automated and the individual steps can be changed if a more fitting method is discovered. Because of automation the process is less biased than surveys that use opinion-based approach.

The clustering enables a researcher to get a quick overview of the topics published in the selected body of literature. The system may reveal unexpected articles under a topic label, because an article can be connected to the cluster via keywords other than the obvious cluster label. Thus, the structural view could be used as a search strategy that complements a simple keyword search. Also, a starting point for a quick literature review on a topic, for example “Security applications of data mining” which was a cluster found in our case study, could be the articles within the particular cluster. Larger clusters corresponding to more general topics, such as “Optimization in data mining”, could be taken as a basis of a new clustering, in order to find and categorize subtopics. For the goals in our case study, though, the initial granularity was sufficient.

Our methodology should be helpful for individuals and companies trying to gain an understanding of large textual datasets, e.g., personal or company internal documentation. It should be useful also for the application field scientists and companies who want to find methods that are currently used widely.

The clustering framework could be used with many different datasets, large or small. There may be scalability issues with larger datasets due to the dimensionality reduction and clustering methods used. Another problem with a large dataset is that some details could be lost in noise. However, when searching for a general overview, this is not a big problem.

Currently the output of our method is a snapshot of current published articles. Combining a longitudinal point of view might reveal long-term trends in research literature. Our approach could benefit from additional information gained from features extracted from abstracts. Abstracts are usually freely available in addition to keywords and titles, whereas other parts of the articles might not be.

Acknowledgements

We are grateful to the anonymous reviewers whose comments have significantly improved the clarity of this paper. This research was partially supported by the Foundation of Nokia Corporation and the Finnish Foundation for Technology Promotion.

References

Agarwal, N., Haque, E., Liu, H., & Parsons, L. (2005). Research paper recommender systems: A subspace clustering approach. In W. Fan, Z. Wu, & J. Yang (Eds.), *Advances in Web-Age Information Management* (pp. 475–491). Berlin Heidelberg: Springer volume 3739 of *Lecture Notes in Computer Science*.

Aljaber, B., Stokes, N., Bailey, J., & Pei, J. (2010). Document clustering of scientific texts using citation contexts. *Information Retrieval*, *13*, 101–131.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*, 351–374.

Bravo-Alcobendas, D., & Sorzano, C. (2009). Clustering of biomedical scientific papers. In *Intelligent Signal Processing, 2009. WISP 2009. IEEE International Symposium on* (pp. 205–209).

Budgen, D., Turner, M., Brereton, P., & Kitchenham, B. (2008). Using mapping studies in software engineering. In *Proceedings of PPIG* (pp. 195–204).

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, *22*, 191–235.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*, 359–377.

Chung, F. R. K. (1997). Spectral graph theory. (p. 2). Providence, RI: AMS Press.

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association: JAMIA*, *13*, 206–219.

Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, *21*, 5–30.

Crimmins, F., Smeaton, A. F., Dkaki, T., & Mothe, J. (1999). Tétrafusion: Information discovery on the internet. *IEEE Intelligent Systems*, *14*, 55–62.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster Analysis*. (4th ed.). London, Arnold; New York: Oxford University Press.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI Magazine*, *17*, 37.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*, 27–34.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*, 471–479.

Glänzel, W. (2003). Bibliometrics as a research field. A course on theory and application of bibliometric indicators. Course Handouts.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Adaptive computation and machine learning. Cambridge, Mass.: MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2011). *The Elements of Statistical Learning*. New York: Springer.

Ivancheva, L. (2008). Scientometrics today: A methodological overview. In *Fourth International Congerence on Webometrics, Informetrics, and Scientometrics & Ninth COLLNET Meeting*.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, *37*, 547–579.

Kitchenham, B. (2004). *Procedures for performing systematic reviews*. Technical Report Keele University and NICTA.

Kriegel, H.-P., Borgwardt, K., Kröger, P., Pryakhin, A., Schubert, M., & Zimek, A. (2007). Future trends in data mining. *Data Mining and Knowledge Discovery*, *15*, 87–97.

Kumar, D., & Bhardwaj, D. (2011). Rise of data mining: Current and future application areas. *IJCSI International Journal of Computer Science Issues*, *8*, 256–260.

Lafon, S., & Lee, A. B. (2006). Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *28*, 1393–1403.

Leydesdorff, L. (2004). Clusters and maps of science journals based on bi-connected graphs in the journal citation reports. *Journal of Documentation*, *60*, 371–427.

Leydesdorff, L., Carley, S., & Rafols, I. (2013). Global maps of science based on the new web-of-science categories. *Scientometrics*, *94*, 589–593.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*, 348–362.

Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O., & O'Blenis, P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association : JAMIA*, *17*, 446–453.

Nadler, B., Lafon, S., Coifman, R., & Kevrekidis, I. G. (2008). Diffusion maps – a probabilistic interpretation for spectral embedding and clustering algorithms. In T. J. Barth, M. Griebel, D. E. Keyes, R. M. Nieminen, D. Roose, T. Schlick, A. N. Gorban, B. Kégl, D. C. Wunsch, & A. Y. Zinovyev (Eds.), *Principal Manifolds for Data Visualization and Dimension Reduction* (pp. 238–260). Berlin Heidelberg: Springer volume 58 of *Lecture Notes in Computational Science and Engineering*.

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, *61*, 1871.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, *314*, 498–502.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*, 265–269.

Szczuka, M., Janusz, A., & Herba, K. (2012). Semantic clustering of scientific articles with use of DBpedia knowledge base. In R. Bembenik, L. Skonieczny, H. RybiÅski, & M. Niezgodka (Eds.), *Intelligent Tools for Building a Scientific Information Platform* (pp. 61–76). Springer volume 390 of *Studies in Computational Intelligence*.

Teregowda, P. B., Councill, I. G., Fernández, R. J. P., Khabsa, M., Zheng, S., & Giles, C. L. (2010). Seersuite: developing a scalable and reliable application framework for building digital libraries by crawling the web. In *Proceedings of the 2010 USENIX conference on Web application development* WebApps'10. USENIX Association.

Tseng, Y.-H., & Tsay, M.-Y. (2013). Journal clustering of library and information science for subfield delineation using the bibliometric analysis toolkit: CATAR. *Scientometrics*, *95*, 503–528.

Venkatadri, M., & Reddy, L. C. (2011). A review on data mining from past to the future. *International Journal of Computer Applications*, *15*, 19–22.

Waltman, L., van Eck, N. J., & Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*, 629–635.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z.-H., Steinbach, M., Hand, D., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*, 1–37.