Noname manuscript No.
(will be inserted by the editor)

Document Clustering of Scientific Texts Using Citation Contexts

Bader Aljaber . Nicola Stokes . James Bailey . Jian Pei

Received: date / Accepted: date

Abstract Document clustering has many important applications in the area of data mining and information retrieval. Many existing document clustering techniques use the "bag-of-words" model to represent the content of a document. However, this representation is only effective for grouping related documents when these documents share a large proportion of lexically equivalent terms. In other words, instances of synonymy between related documents are ignored, which can reduce the effectiveness of applications using a standard full-text document representation. To address this problem, we present a new approach for clustering scientific documents, based on the utilization of citation contexts. A citation context is essentially the text surrounding the reference markers used to refer to other scientific works. We hypothesize that citation contexts will provide relevant synonymous and related vocabulary which will help increase the effectiveness of the bag-of-words representation. In this paper, we investigate the power of these citation-specific word features, and compare them with the original document's textual representation in a document clustering task on two collections of labeled scientific journal papers from two distinct domains: High Energy Physics and Genomics. We also compare these text-based clustering techniques with a link-based clustering

Bader Aljaber
Dept. Computer Science and Software Engineering, The University of Melbourne, Australia
E-mail: baljaber@csse.unimelb.edu.au

Nicola Stokes
School of Computer Science and Informatics, University College Dublin, Ireland
E-mail: nicola.stokes@ucd.ie

James Bailey
NICTA Victoria Laboratory, Dept. Computer Science and Software Engineering, The University of Melbourne, Australia
E-mail: jbailey@csse.unimelb.edu.au

Jian Pei
School of Computing Science, Simon Fraser University, Canada
E-mail: jpei@cs.sfu.ca

algorithm which determines the similarity between documents based on the number of co-citations, that is in-links represented by citing documents and out-links represented by cited documents. Our experimental results indicate that the use of citation contexts, when combined with the vocabulary in the full-text of the document, is a promising alternative means of capturing critical topics covered by journal articles. More specifically, this document representation strategy when used by the clustering algorithm investigated in this paper, outperforms both the full-text clustering approach and the link-based clustering technique on both scientific journal datasets.

Keywords Citation Contexts · Document Clustering · Text Categorization

1 Introduction

The great amount of scientific information being published makes it difficult for users of search engines to identify relevant information. For example, in the biomedical domain alone around 1,800 new papers are published daily [24]. Automatic document clustering provides a possible solution to this information overload problem, whereby users can quickly visualize the search space or search results, using labeled clusters of articles, that have been grouped into topical and sub-topical categories.

Automatic document clustering (that automatically groups related documents into clusters) is a powerful technique for large-scale topic discovery from text that can help to tackle the problem of information overload. For example, document clustering allows unsupervised discovery of the main topics or themes of the documents within a corpus. So, users can scan the clusters to explore documents of interest without having to formulate a query, which is particularly useful when they are unfamiliar with a topic and their exact information need. This is referred to as *clustering based navigation* of the search space. For retrieval, on the other hand, document clustering can be used as a means of improving the efficiency of an IR system by pre-clustering the entire corpus and retrieving clusters rather than documents [48]. This method is said to also improve recall by identifying relevant documents that make no reference to a query term, but which are topically related to other relevant documents in the rank list which do. Liu and Croft in [32] have also used clustering to address term sparsity issues with the document model in a language modelling approach to IR, where the document's containing cluster has been successfully used to interpolate this document model. Furthermore, it can be used as an alternative means of presenting a ranked list of candidate documents in response to a user query[1]. Potentially, this helps users find relevant documents more quickly. Manning et.al. [34] have discussed a number of different clustering applications that take advantage of clustering an IR application.

Traditionally, in document clustering, documents are represented by vectors of term frequencies. Many existing document clustering techniques [55] use this simple "bag-of-words" model to represent a document in a collection. The bag-of-words simply consists of terms that appear in a publication's original source text. Each term is then assigned a "weight of importance" using a weighting metric such as the *tf-idf* weighting scheme [48]. The k-means and hierarchical clustering algorithms are two approaches which use this bag-of-words model.

However, this type of document representation is not always informative for grouping and distinguishing documents, due to two linguistic phenomena: ambiguity and

[1] e.g. As done by the search engine http://clusty.com

synonymy. Ambiguity occurs when documents share lexically similar, but semantically distinct terms (e.g., the money sense of "bank" versus the river sense of "bank"). *Ambiguity* causes errors in text processing tasks, because it can make documents appear more similar than they actually are. *Synonymy*, on the other hand, occurs when semantically related, but lexically dissimilar words make two related documents appear less related than they actually are. In the IR literature the synonymy issue is often referred to as the *vocabulary mismatch problem* [17]. Synonymy has been shown to have a greater effect on IR performance than ambiguity [27]. In this paper, we explore a technique for capturing synonymous and related terms in two scientific domains: High Energy Physics and Genomics. Here are some examples of synonymous terms in these domains:

1. ***Dilaton*** is also known as the ***radion*** or ***graviscalar***.
2. ***Spectrograph*** is equivalent to ***spectroscope***.
3. The unit of measurement ***mole*** is an abbreviated form of the phrase ***gram-molar-weight***
4. ***CRC*** is an acronym used to refer to the disease ***Colorectal cancer***.
5. The ***DLEC1 gene*** (deleted in lung and esophageal cancer 1) can also be referred to by the following synonymous gene names: ***DLC1*** or ***F56***

More specifically, the aim of this paper is to address this issue of synonymy by collecting related and near-synonymous terms from the *citation contexts* of articles for a given journal paper. Citation contexts refer to textual descriptions of a given scientific article found in other articles in the document collection which cite it. Our hypothesis, in this paper, is that these contexts contain useful synonymous and related terms that can be used to boost the accuracy of the similarity calculation between documents in a text clustering application.

We use these citation terms as an alternative representation of an article, which we call the document's *citation representation*. Beside this citation representation, we also present two alternative representations: the standard full-text representation which contains all the vocabulary in the original document, and a hybrid representation which combines original and citation representations together. An additional aim of our experiments is to discover the strengths of these three document representations in the context of three distinct clustering approaches: *Hierarchical Agglomerative Clustering*, *K-means Clustering* and *Bi-clustering*.

Link-based clustering [2,9,26] differs from the other approaches, as it ignores the textual content of the document and instead measures the relatedness of two documents based on the common links (citations) shared by the two documents. Our experiments indicate that such a link-based approach is inferior to a text-based approach for this categorization task.

Our experiments also show that citation contexts can provide relevant synonymous and related vocabulary which helps increase the effectiveness of the bag-of-words representation. More precisely, at the general topic granularity level combining citation terms with the full text representation can significantly improve the document clustering accuracy; when the required topic granularity is more fine-grained, less improvements in document clustering accuracy are observed.

A detailed analysis of our results shows that citation terms tend to capture the general topic keywords of a paper. This lead us to develop an improved approach to standard hierarchical clustering, whereby document similarity is computed using mostly standard full text terms when clusters are small (and topics are thus specific);

while when clusters become large (and topics are thus more general), similarity is then computed mainly using citation terms. We call this method *dynamic* hierarchical clustering, and we show that it can notably outperform the standard hierarchical clustering algorithm on our datasets.

2 Related Work

In this section, we provide a general overview of citation contexts and how they have been used to represent document content in Information Processing applications. We also discuss the similarity between citation contexts and anchor text, which has been used very successfully by the IR community in the area of Web search.

2.1 Citation Contexts

Citations and their use have been of great interest to researchers. One of the seminal works in this area was published by Garfield, who analyzed citation links among scholarly articles [19]. The exploitation of citation links and the context surrounding them, often referred to as citation sentences or *citances*, has also gained much recent attention [35,39]. In these papers, text surrounding a citation is extracted in order to determine the relationship between the two papers connected by that citation, called the *citation function*.

Both [37] and [58] provide a recent review of research surrounding citation analysis. In particular, White [58] states that most of the research in this area is based on a manual analysis of citations from which three major uses of citations are explored: citation categorization where citations are labeled, for example, as conceptual vs. operational, organic vs. perfunctory, evolutionary vs. juxtapositional, and confirmational vs. negative [36]; recurring terms in citances can be used as additional subject headings for indexing purposes; in the context of social networks, citations have been used to explore the citer's motivations (support, oppose or survey) for referring to an earlier related work [37].

There is also some interesting work on citations explored by Nanba et.al. [38,40, 41], where they analyze citations of research papers and automatically classify citation links based on their motivations into three categories, using 160 pre-defined phrase-based rules. The three categories are (i) a comparison to other related papers (either negatively or positively) (ii) building on other related work (iii) others that do not fall into either of the previous two classes. This categorization scheme is used then to build a system for reviewing and survey academic literature.

Work in [37] focuses on the utility of citations in the context of managing the vast of amounts of Life Science literature now available. They identify a number of promising applications of citations in this domain: a source for unannotated comparable corpora, summarization of the target papers, synonym identification and disambiguation, entity recognition and relation extraction, and improved citation indexes for document retrieval.

Teufel and Moens [53] and Siddharthan and Teufel [49] introduced a scientific attribution task, which tries to attribute scientific work to citations. They describe Argumentative Zoning which is a discourse analysis technique that labels sentences, according to their role in the authors argument e.g. contrasting, background. The aim

in this case is to identify the novel claim or contribution of a cited paper by analysing its citations using this technique. Their experiments were conducted on conference articles in computational linguistics and their evaluation, which used comparison to human-annotated attribution, showed a very high agreement (around 80%) with human gold standard annotation.

Another interesting work based on citation contexts is introduced by Elkiss et.al. [16]. They provided a quantitative analysis of the benefits of citation contexts with regards to other applications such as summarization and information retrieval. In particular, they examined the relationship between the abstract and citation contexts of a given scientific paper. Their experiments show that citation contexts may have extra focused information that is not present in an abstract. Therefore, they suggest that citation contexts can be utilized as a different kind of supplementary summary to the traditional abstract.

2.2 Extracting Citation Contexts

An important consideration when using citation contexts is: how to extract them automatically from text? In many cases this is not a straight forward task, since citation maker styles vary from one document to another in the academic literature [43, 45, 54]. For example, there are *formal-textual*, *formal-indexed* and *informal citation* styles. Formal-textual citations use an author-year pair to uniquely identify an entry in the reference list and can be either a syntactic citation (e.g. author-name (year) proposed a method that ...) or a parenthetical citation (e.g. A method proposed by (author-name, year) ...). A formal-indexed citation uses a unique key to refer to a reference in the reference list (e.g. The method introduced in [10] can ...). An informal citation does not require all these pieces of information to distinguish the reference (e.g. author-name has argued that ...).

Many techniques have been proposed to address this problem, such as [4, 5, 43]. A recent attempt to identify and extract citation contexts with high accuracy is introduced in [43]. Powley and Dale collect multiple sources of internal evidence about entities from documents, and integrate citation extraction, reference segmentation, and citation reference matching. In short, they parse the reference list in order to collect entities such as author names and years, and they identify candidate sentences containing these entities. After that, they match reference list items to the candidate sentences using the entities identified earlier. They handle different citation styles and multiple citations in one sentence. Their approach was evaluated with respect to an F-measure and involves author named entity recognition (F = 0.98), citation identification (F = 0.98), and citation reference matching (F = 0.95).

Another interesting and recent work that looks at identifying bibliography items and retrieving citation contexts from a plain text file is introduced by Councill et.al. in [11]. They developed *ParsCit*, which is a system that depends on a machine learning methods coupled with a heuristic processing framework. The system models features useful for identifying bibliography items and matching them with the body text features in order to find relevant citation contexts. The reference list parsing procedure involves a tokenising process based on several metadata fields, such as author and title. For every reference item, one or more regular expressions are produced in order to match the citation contexts in the body of the text. These expressions can handle explicit

citation styles, such as square bracket or parenthetical markers, and implicit citation styles which use the author names and year of publication.

Once the citation markers are identified, determining which terms around the marker actually refer to it, is nontrivial and may even require human interaction. Ritchie et.al. [45] discussed this issue and present some examples of citations where this is the case and proposed methods based on linguistic techniques in order to identify the useful citation terms. Some of their examples show that citations may occur at the start, end, or in the middle of sentences. Other examples show that the sentence boundary can be the boundary of the citation context. In another example, related terms can occur in the following sentences, so the *citation scope* is not at a sentence boundary. Similar arguments can be applied to paragraph and section boundaries. As it is difficult to automatically decide which terms in the citing document reference the cited document, in our work we have extracted contexts around citation references at different window sizes i.e., x terms before and after the citation marker. A more detailed discussion of this method is postponed until Section 3.

2.3 Citation Context use in Ad hoc Retrieval

Many popular literature search engines, such as CiteSeer[2] [29] and Google Scholar[3], also use the links between articles and documents provided by citations to enhance their ranked retrieval results. In both cases, these retrieval systems provide researchers with a means of crawling and navigating through the network of scholarly scientific articles (that is, the citation graph) in a particular domain. Citation links have also been used in those search engines to analyze research trends, and discover the relationships between publications and their ranking in terms of the number of times they have been cited [20].

Bradshaw [7,8] introduced a novel document indexing scheme based on citations called Reference Directed Indexing (RDI). RDI uses terms in citation sentences to index a cited article. Documents are then ranked with respect to the following metrics: the relevance score between document index terms (from the citation sentences) and the query terms, and the number of papers citing that document. Hence, highly cited documents will be ranked higher than documents with lower numbers of citations even if their term indexes have the same number of query terms. The performance of RDI was evaluated against the standard vector-space model which uses *tf-idf* weighting method and the Cosine similarity metric. RDI achieved better precision on the top 10 retrieved documents (statistically significant at 99.5% confidence) [6,8]. In addition, it has been experimentally shown in additional researches [6,45] that good index terms for scholarly IR systems can be found in the documents that cite others.

Similarly, Ritchie et.al. in [46] presented the results of experiments using terms from citations for scientific literature search. They used terms used by citing documents to describe that document, in combination with terms from the document itself. The authors investigated the effect of weighting citation terms differently relative to document terms. In other words, the citation terms are added in duplicate to the document, to achieve the desired weight. Only a small range of weights were tested. Also, they used a range of standard performance measures and t-test for statistical significance

[2] Scientific Literature Digital Library, http://citeseer.ist.psu.edu

[3] Google search engine, for peer-reviewed scholarly literature, http://scholar.google.com

and ran the queries through several standard retrieval models, as implemented in the Lemur Toolkit[4]: Okapi BM25, KL-divergence and Cosine similarity. In each run, 100 documents were retrieved per query. Overall, they found that IR performance is higher with citation terms than without, for all models, for all measures, with the exception of Okapi run. Also, the performance increases as citation terms are weighted more highly.

The difference between works presented in [7,8] and [46] is that the former ones index documents based on the citation terms only, so a document must be cited at least once (by a document available to the indexer) in order to be indexed; whereas the latter one indexes every document based on the combination between citation terms and terms from the document itself. Compared with our work, the authors of those papers have analyzed the performance of information retrieval systems using citation terms, whereas in our paper we are investigating the performance of a document clustering task based on three different representations which will be described with details in Section 3.1. Moreover, we evaluate the use of these three different representations based as a means of capturing the topic granularity of the documents, for two different types of datasets.

In [45], Ritchie et.al. compared the difference between citation terms extracted both manually and automatically (using a fixed window size) from citing articles of a given paper. They also compared these citation terms with the original terms in the paper. Their observations indicated that citation terms could be beneficial in an IR application. However, the effectiveness of a document index enhanced with citation terms was not explored. Similarly, as already stated Bradshaw's [7,8] document index consisted only of citation terms, where the original text of the document was ignored.

Hence, the novelty of the work presented in this paper lies not only in the fact that a new application of citation contexts is presented (cluster generation), but also by the fact that we explore the effectiveness of a combined document representation consisting of both *original document and citation terms*.

2.4 Anchor Text use in Web Retrieval

Another area where link structure analysis has played a critical role is, in the development of web search engines. In the same way that citations infer the importance or relatedness of a scientific works, hypertext links between web pages can provide a measure of content quality and similarity. There are two important algorithms which exploit link structure in this area: *PageRank* which is a query-independent link analysis algorithm [9] and *HITS* which is a query-dependent algorithm and stands for Hyperlink Induced Topic Search[26].

Many researchers in this area have also explored combining web page content with hyperlink information in the clustering of web search results. In [56,57] for instance, a content-link coupled clustering algorithm is introduced, which *linearly* combines text similarity information with link similarity or co-citation similarity information. Their results show, in general, that the average entropy for term-based clustering is higher than the average entropy for link-based clustering; which means that many noisy pages are clustered because they have a high term overlap. Although link-based clustering can reduce this problem, it still suffers from the shortcoming that pages with few inlinks do not have sufficient citation data to create a suitable document representation.

[4] http://www.lemurproject.org/

However, despite these shortcomings with link information, other researchers [23], have observed similar boosts in classification performance when full-text and links document representation strategies are combined with anchor text. *Anchor text* is defined as the text encompassed by a '*<a href*' tag in a HTML document. For instance, in this snippet of text (Google), the word *Google* represents an anchor text snippet. The importance of *extended anchor text* has also been demonstrated [21]. Extended anchor text refers to text surrounding the vocabulary outside of the hypertext link, which is defined by a fixed window size. In addition, researchers have included surrounding *headings* and other *highlighted text fragments* in their extended anchor text definition.

There is a definite parallel between the *anchor text* and *citation contexts* of scientific literature: they both provide a semantic linkage between documents. However, there are also a number of critical differences between them:

1. Anchor text links in web pages are often noisy, as they may be just commercial or navigational links; whereas links of citation contexts are curated and purposefully inserted. We are aware that citation links could also contain some noise. However, generally speaking, literature citations are included by the authors with a specific purpose in mind. For example, when authors cite papers they justify their use with citation contexts that either negatively, positively, or neutrally comment on some related work. Authors also make use of citations to help explain their work and its significance with respect to the related literature. So, literature citations are less likely to be made for no reason. In contrast, web links are commonly inserted, without even the agreement of the authors (e.g. advertisements), or they can be just navigational links with uninformative anchor text such as "click here". Also, anchor text links may be misused in order to influence ranking algorithms such as PageRank, where popular (or even irrelevant) web sites are inserted in order to increase the importance of a page.
2. Links of anchor text are heterogenous; whereas links of citation contexts are homogenous. This means that anchor text links of a given page can link to any kind of object, another web page, a music file, an image; whereas literature citations always link to textual documents (i.e. other publications, reports).
3. Links of Anchor text are dynamic (i.e., the author of web page is able to change them at any time); whereas links of citation contexts are static (i.e., the author of scientific paper is not able to change citations once the paper is published in a journals or proceedings).
4. The window size of extended anchor text is relatively small ($\sim$ 8 words from both sides of the anchor text); whereas the window size of citation contexts is relatively large ($\sim$ 50 words from both sides of the citation marker).

An interesting use of anchor text is presented in [25], which describes an improved version of the *HITS* algorithm [26], where the importance of hypertext links are weighted according to the entropy of the anchor text. The entropy of the anchor text refers to the amount of information the anchor text conveys compared to the actual cited web page. More specifically, this approach attempts to address the issue that most of the content sites usually tend to contain some extra hyper-links, such as navigation panels, advertisements and banners, so as to increase the values of their Web pages in search engines. In other terms, it focuses on improving HITS in order to find informative structures in Web sites. This technique shows better results compared with the HITS algorithm.

In this paper, we explore to what extent citation contexts can improve classification performance. We hypothesize that these descriptive fragments contain synonymous and related terms, that can be used to boost the accuracy of the similarity calculation between documents in a text clustering application for two scientific domains. Although the focus of our paper is text-based clustering algorithms, the success of link-based clustering in both Web IR and the Scientific article search encouraged us to implement a link-based clustering algorithm and compare it against our text-based clustering methods. In the following section, System Description, we describe in more detail the document representation and clustering strategies explored in this paper.

3 System Description

In practice, the bag-of-words model is only effective for discovering the relatedness between documents when these documents share a large proportion of lexically equivalent terms. In other words, instances of synonymy (e.g., the term "physics" and the phrase "physical sciences" are semantically equivalent as defined by *WordNet*) between related documents are ignored, which can reduce the effectiveness of applications using a standard full-text document representation. Consequently, our goal is to discover the benefits that can be gained from a citation representation in the context of a document clustering task, where the domains are High Energy Physics and Genomics. In this paper, we compare the performance of the citation representation against two alternatives, namely an "original" and a "combined" representations. The original representation is a baseline representation, which consists of all the non-stop words mentioned in the original document; the combined representation is a combination of this baseline and the words contained in the citation representation.

The power of these three distinct representations is investigated in the context of three clustering techniques: Hierarchical Agglomerative Clustering (HAC), K-means clustering and Bi-clustering. The remainder of this section, provides additional details on these two system variables (the document representation, and the clustering algorithm).

3.1 Document Representation

Original Term Representation. For a given document, we build a weighted term vector which consists of the most frequent terms mentioned in the original source text. The degree of frequency of terms (a threshold set equal to 3) is specified in order to pick up the frequent terms and to eliminate trivial ones in the document.

All stopwords were removed, and the Porter Stemming Algorithm [42] was applied before the frequency counting was performed, in order to take account of words that only have slight morphological differences (such as plurals). These stopword-removal and stemming processes were also applied when generating the other document representations presented in this section. The *tf-idf* metric is then used to measure the weight of the 'importance' of terms in a document.

Citation Term Representation. The citation term representation for each document is generated from all its citation contexts found in the dataset. More specifically, for every document in both our collections, we automatically extracted all of the citation

sentences that other documents used to refer to it. In Section 2.2, we discuss the difficulty of this task given the *diversity of citation markers* used in the literature, and the added difficulty of detecting the *scope or extend of the citation*. What follows is an explanation of how we dealt with these issues in our work.

One of the major advantages of using our Genomic and Physics datasets is that they already come with annotations that specify which sentences links to what paper. So, the citations in the body of the paper are related to the bibliography items listed in the References section of a publication, by using the HTML anchor tags (e.g. <A HREF=) and LaTeX tags (e.g. \cite {}). The bibliography entries were parsed in order to obtain the unique ID for every document present in the bibliography.

The source of documents with resolved citations was next passed through a set of Java and Perl parsers, that split each document into a format of one sentence per line. During this document parsing, papers with citations were retained. Next all the sentences containing citations were extracted from all the processed documents (and extended from the previous or following sentences to a fixed window size) and grouped into a citation context representing the paper the sentences were citing. If a sentence had citations to more than one paper, it was put into each of the respective citation contexts.

This approach is simplistic, but nevertheless performs well: from a small study of 10 journal documents taken from both our Physics and Genomic data collections, we found and correctly matched (448 out of 466) and (321 out of 330) citations with their corresponding reference (96% and 97%), respectively.

Once the citation markers are identified, the scope of each citation must be determined. As it is difficult to automatically decide which terms in the citing document reference the cited document, we have extracted contexts around citation references at different window sizes. For example, taking 10, 30, 50 terms before and after the citation reference. We also extracted only the citing sentence, regardless of its length.

After conducting a statistical analysis and comparison between these different window sizes, based on the quality and ability of providing related terms to the cited documents, we found that a window size of *50* words from either side of the citation reference generally works well, a finding in agreement with previous work [7,8]. Therefore, in all our reported experiments, we adopt a window size value of 50. The following is an example of a citation context:

The very low-energy Hawking radiation from a massive black hole has non-thermal correlations, which contain detailed information about Planck-scale physics [*]. The phenomenon is reminiscent of the imprinting of planckian fluctuations onto the microwave background radiation by inflation.[5]

Our window size of 50 words from both sides of the citation marker [*] is collected regardless of the sentence boundaries and it must occur within one paragraph. In cases where the citing sentence cites multiple papers, all these cited papers will have the citation context in common. In such cases, where the fixed window size (also called the citation context) comes across another preceding or following citing sentence, the fixed window size will be limited by the sentence boundary (before or after) that citing sentence. Thus, it is guaranteed that no more than one citing sentence is contained

[5] Document number hep-th_9306069 in the physics dataset *arXiv* used for evaluation in this paper

within one citation context. In other words, if there are two citing sentences following each other in one paragraph and their citation contexts are overlapping, the window size of both citation contexts will be reduced (from one side) to address this problem. So, the window size of 50 guarantees that no citation context can have terms from more than one citing sentence (that is no more than one citation context can have terms of a given citing sentence). Otherwise, the sentence boundaries are used to ensure that.

A combination of the Citation and Original Term Representations The third representation consists of words collected from the citation and original representations. Robertson et.al. [47] have analyzed the approach of combining multiple representations to improve the performance of information retrieval systems. The basic idea of the Robertson et.al. scheme is that the structured HTML documents are first transformed into multiple unstructured document representations based on their different fields such as the title and abstract. Every representation-based field is treated separately as a separate collection/index, and assigned a specific weight of importance. For example, relevant documents are retrieved and ranked based on the similarity of their titles (only) with the query terms. Similarly retrieval and ranking is performed on the other text indexes of the other fields. Then all of these ranked lists are combined by linearly combining all of the corresponding similarity scores for each document across each ranked list. This scheme can be useful in the context of multi-field searches, especially when fields are weighted differently according to their importance.

For every document in our experiments, we have merged its citation and original representations into a single representation. There are, in fact, many methods that one could use to perform this merging.

In our experiments, we do something similar to the scheme proposed by Robertson et.al. in [47]. We measure citation terms and original terms weights based on a tf-idf scheme separately and compute the similarity scores for the combined representation of documents using these separate scores. We selectively add only high weighted citation terms (based on a *tf-idf* scheme) into the original representation. More specifically, our methodology for combining the citation and original terms can be explained as follows:

1. After extracting all citation contexts mentioning that document, we remove frequently occurring and basic words in the English language such as *able* and *argument*, according to the list found in the free encyclopedia (that is The Simple English Wikipedia[6]).
2. We calculate the term weights based on *tf-idf* and then select only the top 30% weighted terms.
3. Those selected citation terms are then added to the original terms in order to generate the combined representation. If a term is used both in the original and citation representations, its highest *tf-idf* weight (in either representation) will be used.

3.2 Link-based Clustering Technique Applied

In this paper, we compare text-based clustering approaches against a graph-based clustering technique that was introduced in [13]. This clustering method groups documents

[6] `http://simple.wikipedia.org/wiki/Wikipedia:Basic_English_alphabetical_wordlist`, accessed on May 2008

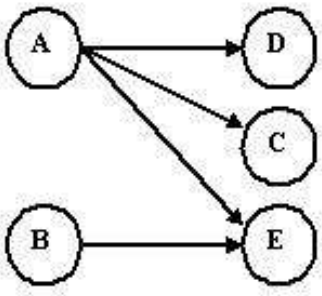

Fig. 1 An example of a survey or review paper.

with respect to their connections to other documents in the citation graph. In a citation graph, nodes represent the documents and the edges represent the citations between documents. The graph clustering technique used here is a multi-level weighted kernel K-means algorithm. Multi-level algorithms repeatedly coarsen the graph level by level until only a small number of nodes are left which are then used to create the initial clustering. Thereafter, the overall graph is un-coarsened level by level, and at each level, the clustering from the previous level is refined using the weighted kernel K-means approach [13].

The technique presented in [13] differs from other multi-level approaches in that it works for a wide class of graph clustering objectives. In other words, it is general, does not constrain clusters to be of equal size, and gives a theoretical guarantee that the refinement step decreases the graph cut objective under consideration [13].

The graph being clustered can be either a weighted or un-weighted graph. In our citation graph, we could treat all citations between documents equally, which would result in an un-weighted graph. Alternatively, we may assign a weight to every citation (link), to obtain a weighted graph. The process of assigning weights might be based on the number of out-links of the citing documents (called fractional citation counting). So, if a paper has ten references in its bibliography, each reference (link) has a fractional weight of 1/10. This technique was previously discussed by Small and Sweeney [51] and it has the generally desirable effect of giving links of papers with short reference lists greater weight, and links of papers with long reference lists, such as review papers, less weight per reference. So, in our citation graph clustering, we have weighed links according to the following equation:

$$LinkWeight = 1/outLinks \tag{1}$$

Where *outLinks* represent the number of references of a paper's bibliography. By using this equation, we can reduce the side effect of survey papers which have so many references in their bibliographies, and tend to cover diverse topics. For example, looking at Figure 1, the link between nodes A and E should be weighted less than the link between nodes B and E. So, based on the previous equation, the weight of the link between nodes A and E will equal (1/3) and the weight of the link between nodes B and E will equal (1). This is due to the fact that node A appears to be a survey paper because of its relative high percentage of references.

3.3 Term-based Clustering Techniques Applied

Text document clustering is an important technique that can be used to automatically organize documents into topically related groups. As already stated, clustering has an important role to play in document search, in particular in the areas of document ranking and results presentation strategies.

As mentioned in the related work section, terms extracted from the original source text (the *bag-of-words* model) are traditionally considered as the standard representation for documents used by many existing applications, including document clustering. Therefore, we aim to discover how much performance improvement can be gained in a document clustering task when citation terms are used by themselves or as supplementary evidence to the *bag-of-words* model.

There are, in fact, many document clustering techniques which can be used. In this paper, we focus on three, namely *Hierarchical* and *K-means* document clustering and *Bi-clustering*.

Hierarchical Document Clustering (HAC). For the three different representations used in this work, we cluster documents using a Hierarchical clustering algorithm. Hierarchical clustering algorithms are either top-down or bottom-up. The top-down clustering algorithm relies on a splitting technique. So, all documents are initially placed in one cluster, the top-down clustering algorithm proceeds by splitting clusters recursively until every cluster contains only one document. Bottom-up algorithms, on the other hand, start with n clusters, where each cluster contains one document. At each step, one determines the closest two clusters using a similarity measure (e.g. single, complete or average linkage methods). The closest two clusters are combined into one cluster. One proceeds until a specified number of clusters are reached or there is only one cluster which contains all documents [34].

In single-linkage clustering, the similarity between two clusters is the similarity of their most similar members. In complete-linkage clustering, the similarity of two clusters is the similarity of their most dissimilar members. The single-linkage (also called the connectedness or minimum method) and complete-linkage (also called the diameter or maximum method) methods merge criteria are local and nonlocal, respectively. Local merge solely considers the area where the two clusters come closest to each other; whereas the distant parts are ignored. Nonlocal merge decisions can be influenced by the entire structure of the clustering [34].

So, intuitively speaking, single-linkage produces chained and skinny clusters and may combine two document topics (clusters), because only two of their members are similar. Complete-link clustering, on the other hand, may enforce two very similar document topics (clusters) not to merge in the early stages, because there is another document topic (cluster), which is less similar and therefore will be grouped first. In average-linkage as the name suggests, the mean distance between clusters is calculated. This can be a good compromise between the extremes of single and complete linkage. In other words, use of an average-linkage method can avoid the problems of single and complete linkage similarities [34].

Therefore, in our methodology, we use the average-linkage method in order to determine the similarity score between two clusters. So, the distance between two clusters is determined based on the average cosine similarity between documents from the first cluster and documents from the second cluster.

K-means Document Clustering. For the three different representations used in this work, we cluster documents using the simple K-means clustering algorithm [22]. K-means clustering is done by minimizing the sum of squares of distances between objects and their corresponding cluster centroid. It groups the objects into K clusters, and keeps iteratively moving the cluster centers and re-assigning objects into clusters, based on minimum distance to the closest cluster's centroid. The process terminates when cluster centers are not moved any more and all objects have been assigned to their closest cluster center. Unlike hierarchical clustering, which groups data objects with a sequence of partitions, K-means (partitional) clustering directly divides data objects into K clusters, without any corresponding hierarchical structure.

Bi-clustering. For the three different representations used in this work, we cluster documents using a bi-clustering algorithm which allows simultaneous clustering of the rows and columns of a matrix [33]. In other words, it is a technique for finding subsets of objects (rows) which exhibit a subset of features (columns) in common. In our scenario, this technique finds subsets of documents (rows) which contain a subset of terms (columns) in common.

Bi-clustering, although more complex, does have some advantages over the hierarchical and K-means clusterings. Hierarchical and K-means clusterings are hard clustering methods, where every document must belong to one and only one cluster, whereas Bi-clustering allows a document to be a member of more than one bi-cluster. This can potentially allow more flexible exploration of topical and sub-topical similarities between documents.

For a detailed explanation of the bi-clustering algorithm used in this work, we refer the reader to Appendix A of this paper.

4 Experimental Methodology and Results

In this section, we describe the experiments we conducted on two document collections:

- *29,555* LaTeX source documents taken from the *arXiv* high energy physics repository used in the KDD Cup 2003 tasks.
- A *162,259* subset of HTML source documents from the HighWire website, which was collected to facilitate the passage retrieval experiments explored at the *TREC* 2006 and 2007 Genomics Tracks.

Regarding the *arXiv* high energy physics repository, it is a large archive of research physics papers used in 2003 for a knowledge discovery and data mining competition held in conjunction with the Ninth Annual ACM SIGKDD Conference[7]. In KDD-2003, there were four varied tasks and every task was a separate competition with its own specific goals. The first task focused on predicting how many citations each paper will receive in the near future. The second task was designed for building a citation graph of a large subset of the archive from only the LaTeX sources. The third task focussed on estimating the popularity of papers based on partial download logs. While the fourth task was an open one, where contestants could devise their own research task, and the most interesting one was determined the winner. Overall, the competition focused on network mining and the analysis of usage logs.

[7] `http://www.cs.cornell.edu/projects/kddcup/`

The TREC Genomics collection, on the other hand, was released to support the evaluation of IR engines who participated in the 2006 and 2007 Genomic track passage retrieval tasks[8]. More specifically, participating systems were required to find exact answer passages to genomic-focussed questions such as "what is the role of PrnP in mad cow disease?", or "how do mutations in the Pes gene affect cell growth?". System performance was evaluated based not only on the precision of the answers retrieved, but also on the extent to which the answer addressed the user's query. The collection was created by obtaining permission to collect papers from 49 publishers who host their publications on the HighWire Press site[9]. TREC organisers crawled the site and after eliminating non-article material ended up a collection containing 162,259 papers.

A reluctance on the part of scientific journal publishers to release journal contents, beyond abstracts, means that both of these collections are special in that they contain full-text documents. The lack of freely available digital journal resources explains the limited amount of reported large scale ad hoc retrieval and clustering experiments, that examine the use of citation contexts in scientific domains. As shown in the Related Work section of this paper, the majority of anchor text/citation context experiments have been performed by the IR Web community, who have easy access to terabytes of Web data.

In our experiments, we work on subsets of *2754* and *3475* documents from the *Physics* and *Genomics* collections respectively. More specifically, we have omitted documents which have been rarely cited by other documents in our collection, as no meaningful citation representations or link structure representations can be built for these documents. We believe that the omission of the documents does not reduce the significance of our results, because our objective here is to study the comparative power of citation features, in comparison to the original bag of words document features. In a 'live' clustering setting, our algorithm would work as follows: if citation information is available, then use it in conjunction with the original document terms, otherwise just use the original document terms.

For the purpose of the experiments reported in this paper, we needed topic labelled scientific articles. In simple terms then, these labels enable us to evaluate the accuracy of our clustering algorithm with respect to the percentage of documents sharing the same label that have been grouped together in the same cluster. Fortunately, the TREC Genomics collection contains labelled articles[10]; however, the KDD Physics collection did not. To address this issue, we mapped the *arXiv* high energy physics papers to their entries in a website[11] that publishes Physics articles for various publishers. Any physical science topic tags found on this site for a given article were assigned to it. Out of 2754 KDD documents we were able to annotate 327 with topic tags[12]. This subset of documents were used in our clustering evaluation. The format and granularity of these medical and physical science labels are discussed in more detail later in this section.

Table 1 shows some statistics for our data collections used. Our datasets used are small by ad hoc retrieval standards. However, they are on the same scale as other collections used in previous research regarding citation contexts. For example, a paper by Elkiss et.al. [16] uses 2497 articles, and two by Ritchie et.al. (in ECIR 2008 [46]

[8] http://ir.ohsu.edu/genomics/

[9] `www.highwire.org`

[10] http://www.csse.unimelb.edu.au/∼baljaber/genomicsIJIR.dat

[11] American Institute of Physics, http://www.aip.org/pacs, the list of 2006

[12] http://www.csse.unimelb.edu.au/∼baljaber/physicsIJIR.dat

Table 1 Table showing some statistics on our Physics and Genomic document collections.

Collection name	Original size of collection	Collection size after citation analysis was performed	Number of labelled documents in subset that were used in evaluation
Physics	29,555	2754	327 for all PACS tag levels
Genomic	162,259	3475	3475, 3470, 3432, 3416 for MeSH term level 1, 2, 3, 4 respectively

and CIKM 2008 [44]) use around 3300 articles. All three of these papers used a single dataset, whereas we have used two datasets, from different domains.

Our experiments were done using a Solaris 9/x86 machine, CPU speed of 3.0GHz and 4 GB of memory. The process of extracting the citation contexts from both dataset and representing documents took around 70 hours. The HAC clustering required around 14 hours. The dynamic HAC and bi-clustering (for all representations) were more computationally intensive and required approx. 70 hours each.

Our implementation of all these methods was not optimised, since time efficiency was not a focus of the investigation. Rather, our sole focus was the investigation of the effectiveness/accuracy of document clustering using citation terms. It is worth noting also that these computation costs are essentially static costs for pre-processing the collection and constructing the clustering. They are not costs which would be incurred at "search time". Also, there are many existing methods that might be used to address the well-known $O(N^2)$ complexity of HAC clustering.

The remainder of this section is ordered as follows: next we partially motivate our hypothesis that citation contexts contain synonymous and related terms, by comparing their overlap with the original text of their corresponding documents in our collections; we then define our quality metric for evaluating the accuracy of our clustering algorithms and document representations on our collections of labelled documents. The objective of these experiments is two-fold: to determine to what extent citation sentences are an appropriate document representation for clustering scientific documents; and to establish which of the proposed clustering algorithms perform best for this task. The section ends with an deeper analysis of our results, and the proposal of a novel dynamic hierarchy (HAC) clustering algorithm.

4.1 How distinctive are the terms in a citation representation?

Before presenting the categorization accuracy of clustering methods, we first provide results on a preliminary experiment that helps to motivate the use of citation representations. Our hypothesis states that citation contexts contain important related and synonymous words that, while being lexically dissimilar to terms in the original cited document, may prove useful for uncovering additional links with topically related documents in the scientific literature. To observe the extent to which citation representations differ from their original documents, we present the following experiment which calculates the cosine overlap between each of the terms in each document and the terms in the citation contexts that refer to that document. Figure 2 presents a histogram of

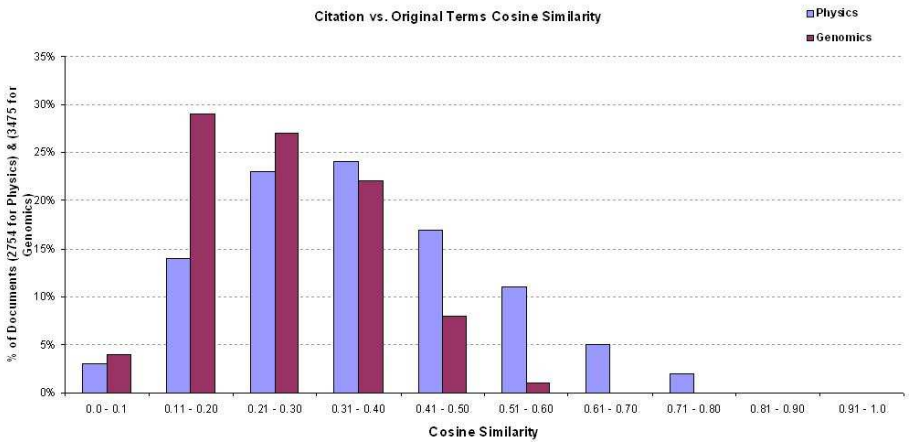

Fig. 2 Graph showing the distribution of the cosine similarity scores of each Original (Full-text) Representation and its corresponding Citation Representation for all selected papers (that have a citation representation) in our Physics and Genomics Collections.

document percentages that fall into different cosine similarity ranges. This graph shows that the majority of documents exhibit low cosine similarity scores with their citation representations. More specifically, the majority of document-citation similarity scores are less than 0.4 for the Physics collection and 0.2 for the Genomics collection. This result is encouraging, as it indicates that citation representations can provide novel vocabulary that could be used to supplement a standard document representation. In the following subsections, we present results that specifically compare and contrast the effectiveness of citation vocabulary in a document clustering application. In particular, results presented in the next subsection will clarify for us whether these citation representation terms, regardless of their uniqueness, are actually relevant and useful for expanding the original context of documents in our collection.

4.2 Clustering Accuracy Results

In this subsection, we present the accuracy of our clustering experiments in a text classification task. We begin with the introduction of our evaluation metric, and the performance of our clustering algorithms and document representations on the Physics dataset outlined earlier in this section. Accuracy results are then presented for the Genomics collection, followed by a deeper manual analysis, and the proposal of a novel dynamic HAC algorithm which utilizes these observations.

As already stated, our evaluation here focuses on calculating the accuracy of the generated clusters. That is, given a set of predefined classification labels, we estimate accuracy based on the number of shared labels among documents in our generated clusters. In this case, documents in our Physics dataset are labelled with *PACS* (Physics and Astronomy Classification Scheme) tags. PACS tags are category and subject descriptors chosen from a controlled vocabulary by the author (or journal editor) to describe the overall topic of the paper. These labels are hierarchical in nature, with general descriptors being at the top of the hierarchy and more specific descriptors at

the lower levels. Obviously if documents in a cluster share all of their *PACS* tags, then we can say that these clusters are highly accurate; an average of these scores then gives us an estimate of the accuracy of the clustering. *PACS* tags were formulated by the American Institute of Physics[13] in collaboration with ICSTI (International Council on Scientific and Technical Information). The *PACS* tag consists of the conjunction of two-digit numbers, decimal point, and two-digit characters. Here are some sample *PACS* tags with their corresponding label descriptors: *63.20.Dj*: *Phonon states and bands, normal, modes, and phonon dispersion*; *63.20.Ls*: *Phonon interactions with other quasiparticles; 63.20.Mt: Phonon-defect interactions.*

To evaluate the accuracy of the clusterings at each PACS level of the hierarchy, we use the Break-Even Point (BEP) metric, where BEP refers to the point at which Precision = Recall. However, since it is time consuming to work out the exact value of the BEP, it is customary to estimate it using the arithmetic mean of Precision and Recall. BEP has been used by many other researchers who have evaluated clustering techniques in the context of a text classification task [1,3,10,15,18,50].

The BEP metric is calculated as follows: for each cluster we calculate a BEP value for each label. Then the cluster with the highest BEP score for a given label is assigned that label. All non-labeled clusters are then ignored. The *macro-average* of these final BEP values is then reported for the clustering. In this case, Precision and Recall for a given label A are defined as follows:

- Precision for a given cluster is defined as the total number of documents in the cluster that are referenced by this label A, divided by the total number of unique labels assigned to documents in the cluster.
- Recall for a given cluster is defined as the total number of documents in the cluster that are referenced by this label A, divided by the total number of documents in the entire collection with label A.

4.2.1 Physics Journal Clustering Results

In our Physics Journal clustering experiments, we evaluated clusters at PACS level 1, 2, 3, 4 and we clustered the dataset into 3, 7, 15, 29 document clusters, respectively. Originally these PACS levels had many more labels; however, we removed all labels which had less than 10 documents each, thus ensuring that all labels, for example with one assigned document, didn't skew the BEP average with 100% precision values[14]

As already stated, different PACS levels have different numbers of labels. Choosing the appropriate number of document clusters without any prior knowledge of the data is a model selection problem which is beyond the scope of this paper. Hence, we set the number of document clusters (the cluster limit) to be identical to the number of "real" categories at each level of the PACS hierarchy. BEP scores for each clustering algorithm on each of our 4 PACS levels are shown in Figures 3, 4 and 5.

We can observe clearly in Figures 3 and 5 that the term-based clustering methods using the 'combined' representation, which is a combination of the 'original' and

[13] American Institute of Physics, http://www.aip.org/pacs, the list of 2006

[14] If label A has only 1 document and this document belongs to cluster C, then cluster C will have a recall value of 100% for this label. Precision on the other hand will vary depending on how many other documents this cluster contains. The BEP score (the average of these two values) could, in this case, look more impressive than real system performance would suggest. This explanation shows how low density labels can incorrectly boost our precision values for this clustering task. Hence, we removed these labels from our evaluation.

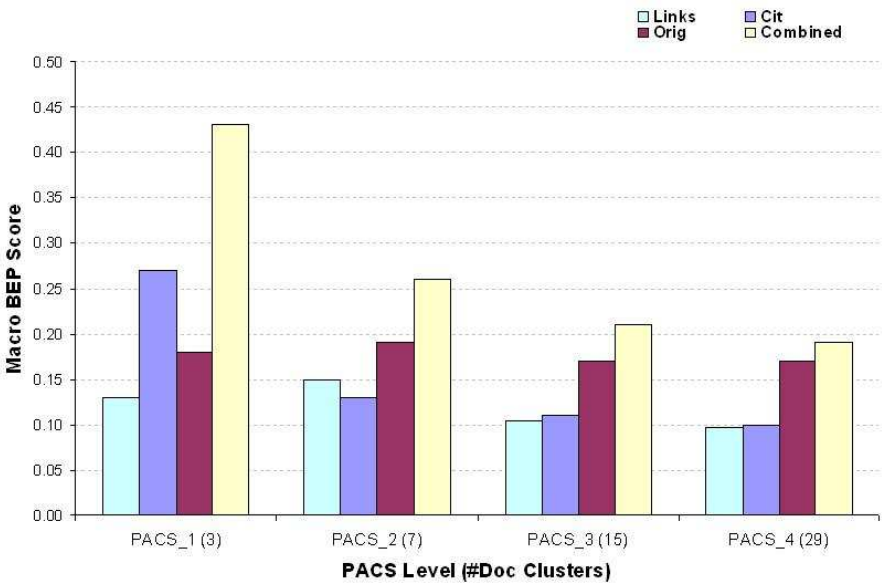

Fig. 3 HAC BEP scores based on PACS labels in the Physics collection.

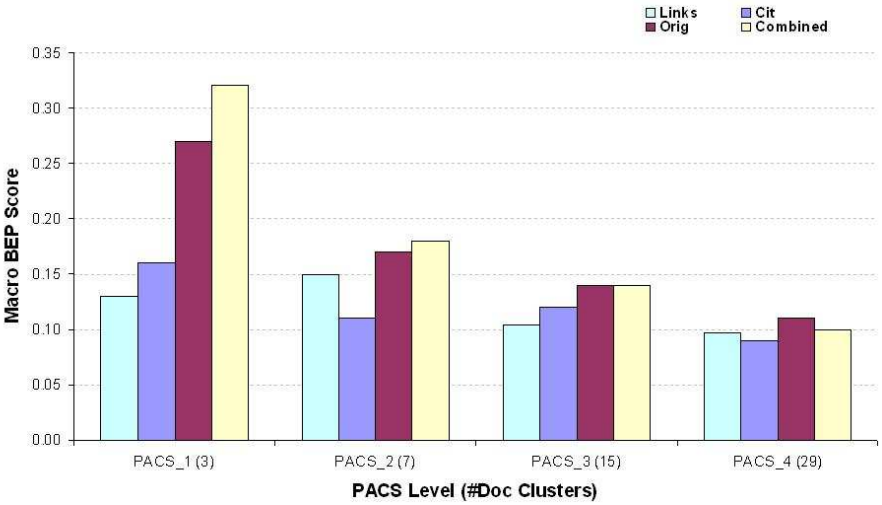

Fig. 4 K-means clustering BEP scores based on PACS labels in the Physics collection.

'citation' representations, tends to outperform both the original representation (the standard full-text document representation) and the links-based clustering method. Figure 4 tells a similar story; however, K-means clustering using the combined representation only outperform others representations on PACS levels 1 and 2. For more specific topic classification (that is PACS level 3 and 4) the term-based clustering technique of K-means using the original representation gives equal if not better results than those using the combined representation.

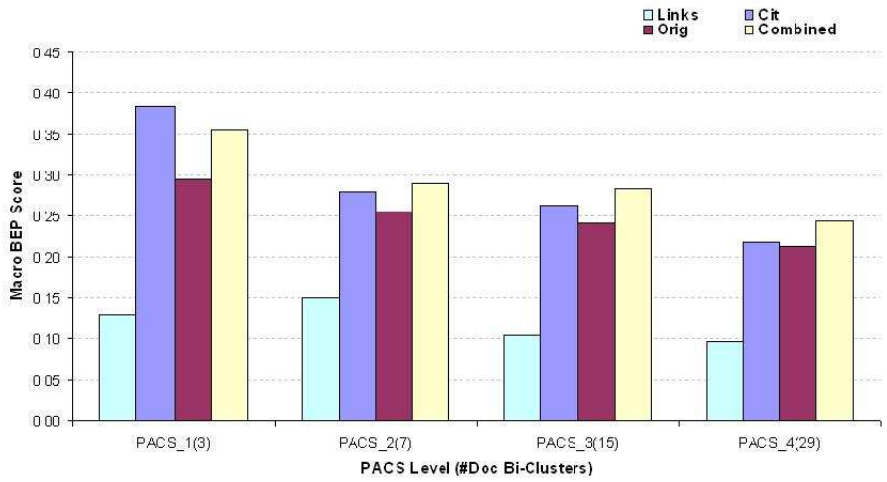

Fig. 5 Bi-Clustering BEP scores based on PACS labels in the Physics collection.

It is worth noting that despite the limited vocabulary size of a citation representation, it still performs competitively against the larger original and combined representations (the citation representation has on average of 108 distinct terms per document compared with 213 for the original representation). This result suggests that when clustering efficiency is of particular concern, a purely citation representation may be a suitable alternative.

With respect to the effectiveness of our clustering algorithms on the Physics dataset, we can see that BEP values for the bi-clustering algorithm are in general higher than for our hierarchical and K-means clustering algorithms. Table 2 shows the average of all four PACS level BEP scores for the best performing representation run of each of our three clustering algorithms. We performed a number of statistical significance tests using the Wilcoxon sign-rank test, and found that the combined representation significantly outperforms the other representations in the HAC and bi-clustering runs (denoted by ‡). K-means using the combined representation also shows improvements over the other representations; however, this increase in BEP was not found to be a statistically significant improvement.

4.2.2 Genomic Journal Clustering Results

In our second set of experiments, performed on a collection of journal papers in the Genomic domain, we used the same evaluation process which focuses on the accuracy of the generated clusters, based on their ability to cluster documents that share topic labels. In this instance, our labels are referred to as *MeSH terms*, where MeSH stands for ***Me***dical ***S***ubject ***H***eadings. Again these labels are part of a huge controlled vocabulary of topic terms used for indexing journal articles and books in the Life Sciences arena. MeSH terms are managed by the United States National Library of Medicine (NLM)[15].

Like PACS tags, MeSH terms are hierarchical in nature, with general descriptors organised at the top of the hierarchy, and more specific descriptors situated at lower

[15] http://www.nlm.nih.gov

Table 2 Table showing the average of all four PACS level BEP scores for the best performing representation run of each clustering algorithm on the Physics collection.

	HAC Clustering Algorithm		
Links	Cite	Orig	Combined
0.121	0.152	0.177	**0.275**‡

	K-means Clustering Algorithm		
Links	Cite	Orig	Combined
0.121	0.120	0.177	**0.180**

	Bi-Clustering Algorithm		
Links	Cite	Orig	Combined
0.121	0.286	0.252	**0.295**‡

‡= Statistically Significant

levels. The *MeSH* term consists of a main heading and some qualifiers. For example, the MeSH term, *Muscle, Smooth/metabolism/*physiology*, consists of a main heading, *Muscle, Smooth*, and a set of qualifiers, *metabolism/*physiology*. In our Genomics collection, every document has been assigned multiple MeSH terms, which amounts to thousands of unique terms. More specifically, there are almost 5,892 distinct main headings, and 14,301 distinct MeSH terms (main headings and qualifiers). Since our evaluation metric, BEP, is suitable only when we have a "reasonable" number of labels, we have simplified all MeSH terms by considering the main headings only. For instance, for *Muscle, Smooth/metabolism/*physiology*, we have only considered its main heading which is *Muscle, Smooth*.

Moreover, we have investigated the power of the citation terms at different topic granularities. So, for topic granularity level 1, we have simplified all MeSH terms (main headings) by mapping them to their most general concept node, that is, to one of the MeSH *basic descriptor* tags. For example, *Medicine, Arabic* has the following id: *E02.190.488.510*, where E02 is a basic descriptor that references the concept *Therapeutics*. Similarly, we have also conducted our evaluation based on the other topic granularity levels. So, looking at the previous MeSH example, for topic granularity level 2, 3 and 4, we have used *E02.190* (that references the concept *Complementary Therapies*), *E02.190.488* (that references the concept *Medicine, Traditional*) and *E02.190.488.510* (that references the concept *Medicine, Arabic*) respectively. Different topic granularity levels have different numbers of labels (categories) and when we evaluate our clusterings based on those topic granularity levels, we need to set the number of document clusters (the cluster limit) to be identical to the number of labels at each level. Thus, for level 1, 2, 3 and 4, we clustered the dataset into 98, 367, 349 and 539 document clusters respectively.

Looking at Table 3, at topic granularity level 1, we can see that the Combined representation based clusters are consistently better than the other representation runs, and the K-means and Bi-clustering results are statistically significantly better than the other representations (denoted by ‡). However, at topic granularity level 2, 3 and 4, we can see that sometimes citation terms when combined with the full-text content of documents brings a slight (but not significant) improvement.

In contrast to our Physics collection experiments, the citation representation runs are consistently poorer performing than the original representation runs. Overall BEP

Table 3 Table showing our evaluation metric based on BEP values for our HAC, K-means and Bi-clustering algorithms conducted on TREC Genomics collection.

		HAC Clustering Algorithm		
Topic Granularity (#Doc Clusters)	Links	Cite	Orig	Combined
level 1 (98)	0.207	0.332	0.362	**0.373**
level 2 (367)	0.203	0.275	**0.332**	0.326
level 3 (349)	0.172	0.226	**0.273**	0.268
level 4 (539)	0.198	0.239	0.263	**0.264**

		K-means Clustering Algorithm		
Topic Granularity (#Doc Clusters)	Links	Cite	Orig	Combined
level 1 (98)	0.207	0.368	0.390	**0.412**‡
level 2 (367)	0.203	0.358	0.384	**0.385**
level 3 (349)	0.172	0.331	0.351	**0.357**
level 4 (539)	0.198	0.330	**0.348**	0.336

		Bi-Clustering Algorithm		
Topic Granularity (#Doc Bi-Clusters)	Links	Cite	Orig	Combined
level 1 (98)	0.207	0.202	0.254	**0.282**‡
level 2 (367)	0.203	0.126	**0.228**	0.225
level 3 (349)	0.172	0.105	**0.199**	**0.199**
level 4 (539)	0.198	0.105	**0.205**	0.201

‡= Statistically Significant

scores are higher on the Genomic collection than they were on the Physics clustering experiments. The relative performance rankings of the clustering algorithms are also different on this dataset. K-means is our best performing algorithm, followed by HAC and Bi-clustering; whereas previously, the Bi-clustering approach was our best performer and K-means was our weakest.

4.2.3 Manual Analysis of Clustering Results

Our BEP evaluation results indicate that citation terms are most effective when used as supplementary evidence of document content with the original document representation (that is, the *combined* representation). Looking for additional trends in our results, we can see that the performance of the combined representation decreases as *topic specificity* increases for all clustering algorithms (see Figures 3, 4 and 5) on the Physics dataset and (see Table 3) on the Genomic dataset.

We also performed a manual evaluation of our results to try and explain this trend, and found that in many cases, citation terms describe general aspects of the topic of an article, because the citing author usually wants to save space by referring the reader to the original source paper for additional information. Here are two citation sentence samples exhibiting this characteristic:

- *A "stretched horizon" [*] can be defined as the place at which modes asymptotically have energies equal to the Hawking temperature are blueshifted to the Planck scale.*[16]
- *The XRCC1 194Arg and 399Gln alleles were associated with increased risk for oral cavity and pharyngeal cancers [*].*[17]

[16] Document number hep-th_0209231 in the Physics *arXiv* dataset.
[17] Document number 11375899 in the TREC Genomics dataset.

Table 4 Table showing the evaluation metric based on BEP values for our Dynamic HAC Document Clustering algorithm conducted on TREC Genomics collection.

		HAC Clustering Algorithm			Dynamic
Topic Granularity (#Doc Clusters)	Links	Cit	Orig	Combined	Combined
level 1 (98)	0.207	0.332	0.362	0.373	**0.381**‡
level 2 (367)	0.203	0.275	**0.332**	0.326	0.326
level 3 (349)	0.172	0.226	**0.273**	0.268	0.267
level 4 (539)	0.198	0.239	0.263	0.264	**0.267**

‡= Statistically Significant

The first citation mentions only the general definition of a "stretched horizon" proposed by the cited paper. Hence, readers are required to go back to the cited paper for more technical details. Similarly, the second citation does not provide any specifics on the biological reason for the pathogenic outcome caused by mutations in the XRCC1 gene.

4.3 Dynamic Hierarchical Clustering

The above observation regarding citation term's tendency to capture the general topic keywords of a paper, suggests an improved approach to hierarchical clustering may be possible. More specifically, in the early stages of hierarchical clustering, it may be better to compute document similarity mainly on the original term overlap between documents, since each cluster is very small and thus specific. In the latter stages of hierarchical clustering, it may be better to compute document similarity mainly on the citation term overlap, because clusters are much larger at that stage and thus more general.

To achieve this type of effect, we consider making the behavior of the hierarchical clustering algorithm more similar to the notion of the combined representation, that was described earlier. More precisely, at the beginning of the hierarchical clustering algorithm, where clusters represent specific topics, we add just a small proportion of terms from the top weighted *tf-idf* citation terms to the original representation, in order to build the combined representation. At the end of the hierarchal clustering process, where clusters represent general topics, we add a larger proportion of top weighted citation terms to build the combined representation.

More specifically, our *dynamic* hierarchal clustering algorithm divides the hierarchy into 10 equal parts. In the first part (at the beginning of the hierarchal clustering algorithm), the combined representation consists of only the original terms; in the second part, the combined representation consists of all original terms plus the top 10% of the top weighted citation terms. So, as the algorithm approaches the root of the hierarchy, more and more citation terms are added. Figure 6 and Table 4 show the results achieved by this technique on the Physics and Genomics collections, respectively. On the Physics collection, dynamic HAC clustering results are better than the static HAC clustering at PACS level 2 and 3. On the Genomics collection, dynamic HAC clustering achieves more significant performance improvements (denoted by ‡) only at topic granularity level 1. Finally, this dynamic approach provides strong evidence that using a more complex term selection strategy can pay off when using citation terms in a clustering application.

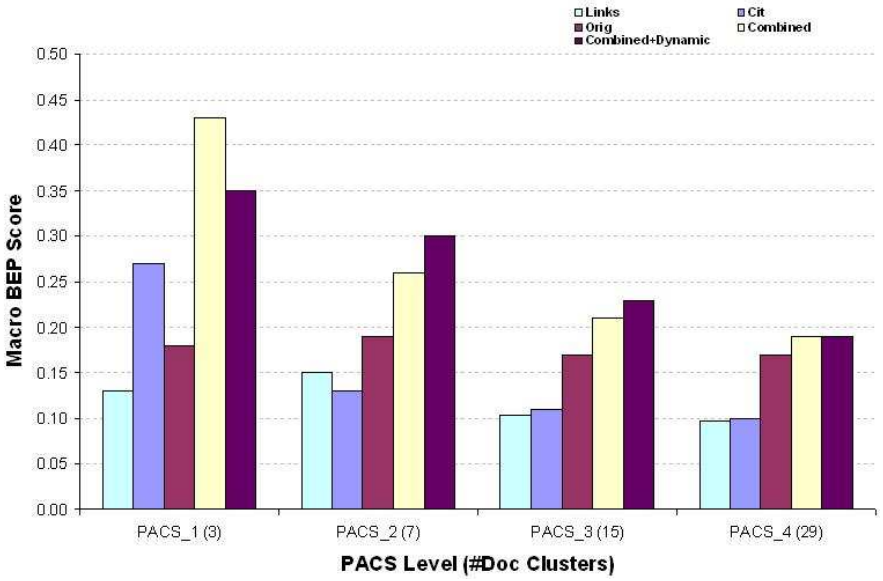

Fig. 6 Dynamic HAC Document Clustering Evaluation for Physics collection based on PACS tags

The traditional HAC algorithm requires the calculation of all distances between all pairs $O(N^2)$. In our dynamic HAC, we change the representations in stages throughout the clustering process, so we are introducing inversions into the cluster hierarchy. As a result, the distance calculation process is repeated every time we change the representations. Although this technique will affect the efficiency of the clustering process, the final results can be more accurate. By presenting this dynamic HAC we are not arguing that it is efficient, but rather it can generate a more accurate clustering. In future work, we plan to extend our dynamic HAC algorithm by exploring the use of an approximate HAC algorithm[28], which will reduce clustering time by careful choosing a subset of distances to calculate, thus avoiding the calculation of all pairwise distances.

5 Discussion

In summary then, the following important conclusions can be drawn from the experiments presented in this paper:

- Using citation terms can improve document clustering accuracy, when combined with a traditional full-text representation of the document. This improvement becomes significant when documents are characterised at a general (rather than a specific) level of topic granularity.
- Document representations consisting of only citation terms are surprisingly effective in a clustering application. This result indicates that citation representations, which typically contain less terms than full-text representations, can improve runtime efficiency without causing a critical drop in clustering accuracy.
- While our Bi-clustering approach performed best on the Physics collection, and K-means clustering outperformed all other approaches on the Genomics collection,

the HAC clustering algorithm exhibited the most stable performance across both datasets.
- In general, citation terms tend to capture general topic keywords rather than specific ones. This observation leads to our proposal for a modified HAC approach that uses different mixes of citation and original terms for the similarity computation, for each level of the hierarchy. Our results show that this approach is a promising dynamic clustering solution.
- Finally, our results show that a link-based clustering approach, which use co-citation information to determine document similarity, is inferior to a text-based clustering approach on this text categorization task.

Regarding taking the Top 30% (based on tf-idf) from the citation context when we build the combined representation, this parameter is based on experiments which we have conducted. We found that choosing the top 30% gives good results, although varying this number or even taking 100% of the citation terms still gives quite acceptable results. This is confirmed in Figure 7 which shows how HAC clustering performance (on our TREC genomics dataset) varies with term percentage.

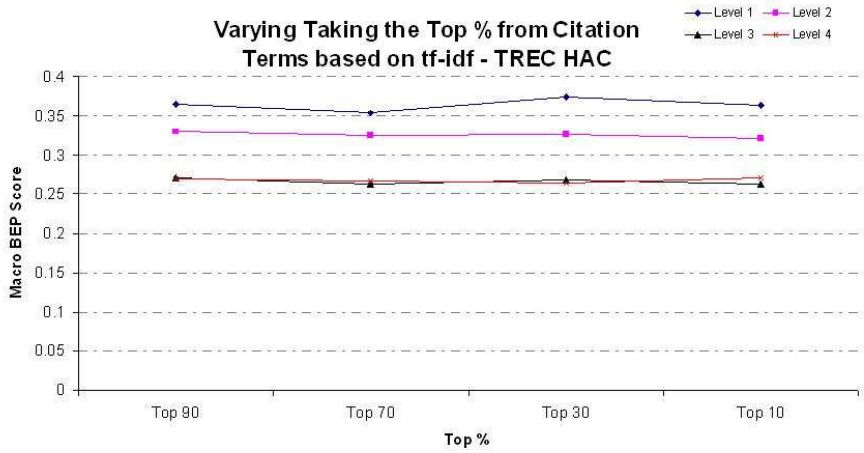

Fig. 7 Change in HAC performance as the percentage of top citation terms combined with the original representation is varied.

One of our intentions for future work is to explore alternative term weighting schemes for our citation terms. More specifically, work by [23] in the Web IR community has shown that *distance weighting* can address the problems associated with the use of a fixed window size for capturing citation contexts. More specifically, windowing strategies often capture erroneous terms which are outside the scope of influence of the current citation. A term weighting scheme that factors in the distance of the term from the anchor (or citation, in our case) can, according to [23], significantly improve results in an IR setting. Similar gains may be possible in this application domain also.

6 Conclusions

To conclude, our results show that citation sentences are a good alternative document representation, which provide additional topical information on the source document. In particular, they may contain useful synonymous and related terms that can boost the accuracy of the similarity calculation, which is an integral part of applications such as IR and document clustering. We did not explicitly analyse the content of these citation representations to assert this hypothesis. Instead we set up a series of experiments to determine how effective a citation representation is when compared with an original full-text representation of a document in a clustering application.

Overall, our results show that the combination of the original and the citation representations is the most effective means of capturing the content of the scientific documents in our collection. In other words, the citation representation should not be used to replace the original full-text version, unless efficiency is of particular concern to the application.

Another important contribution of this work, is the analysis and use of referential contexts within the domain of scientific publications. Much of the previous work in this area has focussed on anchor text use in Web retrieval and clustering applications. During the course of this work, we have also developed two new test collections of labelled scientific documents in two domains: Genomics and Physics, which will help to facilitate future research in this area.

References

1. K. Aas and L. Eikvil. Text categorisation: a survey. *Technical Report NR 941, Norwegian Computing Center*, June 1999.
2. R. Angelova and S. Siersdorfer. A neighborhood-based approach for clustering of linked document collecitons. In *Proceedings of the 15th ACM conference on Information and knowledge management*, pages 778–779, 2006.
3. R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text categorization. *J. Mach. Learn. Res.*, 3:1183–1208, 2003.
4. D. Bergmark. Automatic extraction of reference linking information from online documents. Technical Report CSTR 2000-1821, Cornell Digital Library Research Group, 2000.
5. D. Bergmark, P. Phempoonpanich, and S. Zhao. Scraping the ACM digital library. *SIGIR Forum*, 35(2):1–7, 2001.
6. S. Bradshaw. Document indexing vocabularies: Reference vs content. *Northwestern University (Technical Report, NWU-CS-01-7)*, 2001.
7. S. Bradshaw. Reference directed indexing: Indexing scientific literature in the context of its use. *Ph.D. dissertation, Northwestern University (Tech Report NWU-CS-02-7)*, 2002.
8. S. Bradshaw. Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries*, pages 499–510, 2003.
9. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh International Conference on World Wide Web*, pages 107–117, 1998.
10. F. Chik, R. Luk, and K. Chung. Text categorization based on subtopic clusters. *Natural Language Processing and Information Systems*, 3513:203–214, May 2005.
11. I. G. Councill, C. L. Giles, and M. Y. Kan. Parscit: An open-source crf reference string parsing package. In *Proceedings of Language Resources and Evaluation Conference (LREC 08)*, 2008.
12. M. Dash and H. Liu. Feature selection for clustering. In *Proceedings of The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 110–121, 2000.
13. I. Dhillon, Y. Guan, and B. Kulis. Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(11):1944–1957, November 2007.

14. I. Dhillon, J. Kogan, and M. Nicholas. Feature selection and document clustering. survey of text mining. pages 73–100. Springer-Verlag, 2004.
15. S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the ACM Conference on Information and Knowledge Management*, pages 148–155, 1998.
16. A. Elkiss, S. Shen, A. Fader, G. Erkan, D. J. States, and D. R. Radev. Blind men and elephants: What do citation summaries tell us about a research article? *JASIST*, 59(1):51–62, 2008.
17. G. Furnas, T. Landauer, L. Gomez, and S. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.
18. E. Gabrilovich and S. Markovitch. Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*, pages 1301–1306, 2006.
19. E. Garfield. Science citation index, a new dimension in indexing. *Science*, 144(3619):649–654, 1964.
20. C. Giles, K. Bollacker, and S. Lawrence. Citeseer: an automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98, June 1998.
21. E. Glover, K. Tsioutsiouliklis, S. Lawrence, D. Pennock, and G. Flake. Using web structure for classifying and describing web pages. In *Proceedings of the World Wide Web Conference*, pages 562–569, 2002.
22. J. Hartigan and M. Wong. A k-means clustering algorithm. In *Applied Statistics, 28*, pages 100–108, 1979.
23. T. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the web. In *Proceedings of the World Wide Web Conference*, pages 432–442, 2002.
24. L. Hunter and K. Cohen. Biomedical language processing: What's beyond pubmed? *Mol Cell*, 21(5):589–594, March 2006.
25. H.-Y. Kao, M.-S. Chen, S.-H. Lin, and J.-M. Ho. Entropy-based link analysis for mining web informative structures. In *Proceedings of the ACM Conference on Information and Knowledge Management*, pages 574–581, 2002.
26. J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
27. R. Krovetz and W. Croft. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141, 1992.
28. M. Kull and J. Vilo. Fast approximate hierarchical clustering using similarity heuristics. *BioData Mining*, 9(1), 2008.
29. S. Lawrence, C. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.
30. J. Liu, S. Paulsen, X. Sun, W. Wang, A. Nobel, and J. Prins. Mining approximate frequent itemsets in the presence of noise: algorithm and analysis. In *Proceedings of the 6th SIAM International Conference on Data Mining (SDM)*, pages 405–416, 2006.
31. T. Liu, S. Liu, Z. Chen, and W. Ma. An evaluation on feature selection for text clustering. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, pages 488–495. Washington DC, 2003.
32. X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK*, pages 186–193, 2004.
33. S. Madeira and A. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.
34. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008.
35. R. Mercer and C. D. Marco. A design methodology for a biomedical literature indexing tool using the rhetoric of science. In *Proceedings of the BioLink workshop in conjunction with Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL)*, pages 77–84, 2004.
36. M. Moravcsik and P. Murugesan. Some results on the function and quality of citations. *Social Studies of Science*, 5:86–92, 1975.
37. P. Nakov, A. Schwartz, and M. Hearst. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*, pages 81–88, 2004.

38. H. Nanba, T. Abekawa, M. Okumura, and S. Saito. Bilingual presri integration of multiple research paper databases. In *Proceedings of RIAO*, pages 195–211, 2004.
39. H. Nanba, N. Kando, and M. Okumura. Towards multi paper summarization using reference information. In *Proceedings of the 16th International Joint Conferences on Artificial Intelligence (IJCAI-99)*, pages 926–931, 1999.
40. H. Nanba, N. Kando, and M. Okumura. Classification of research papers using citation links and citation types: Towards automatic review article generation. In *The 11th SIG Classification Research Workshop, Classification for User Support and Learning, 2000.11, in Chicago, USA*, pages 117–134. The American Society for Information Science (ASIS), 2000.
41. H. Nanba and M. Okumura. Automatic detection of survey articles. In A. Rauber, S. Christodoulakis, and A. M. Tjoa, editors, *Research and Advanced Technology for Digital Libraries, 9th European Conference, ECDL 2005, Vienna, Austria, September 18-23, 2005, Proceedings*, volume 3652 of *Lecture Notes in Computer Science*, pages 391–401. Springer, 2005.
42. M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
43. B. Powley and R. Dale. Evidence-based information extraction for high-accuracy citation extraction and author name recognition. In *Proceedings of the 8th RIAO International Conference on Large-Scale Semantic Access to Content*, 2007.
44. A. Ritchie, S. Robertson, and S. Teufel. Comparing citation contexts for information retrieval. In J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K.-S. Choi, and A. Chowdhury, editors, *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, pages 213–222. ACM, 2008.
45. A. Ritchie, S. Teufel, and S. Robertson. How to find better index terms through citations. In *Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval?*, pages 25–32. Sydney, 2006.
46. A. Ritchie, S. Teufel, and S. Robertson. Using terms from citations for information retrieval: Some first results. In *Proceedings of the 30th European Conference on Information Retrieval (ECIR)*, pages 211–221, 2008.
47. S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM)*, pages 42–49, New York, NY, USA, 2004. ACM.
48. G. Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.
49. A. Siddharthan and S. Teufel. Whose idea was this and why does it matter? attributing scientific work to citations. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 316–323, 2007.
50. N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the international ACM SIGIR conference on Research and development in information retrieval*, pages 208–215, 2000.
51. H. Small and E. Sweeney. Clustering the science citation index using co-citations. *Scientrometrics*, 7(3-6):391–409, 1985.
52. B. Tang, M. Shepherd, E. Milios, and M. Heywood. Comparing and combining dimension reduction techniques for efficient test clustering. In *Proceedings of the Workshop on Feature Selection for Data Mining, SIAM International Conference on Data Mining (SDM)*, pages 17–26, 2005.
53. S. Teufel and M. Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.
54. S. Teufel, A. Siddharthan, and D. Tidhar. Automatic classification of citation function. In *Proceedings of EMNLP-06*, 2006.
55. E. Voorhees. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Inf. Process. Manage.*, 22(6):465–476, 1986.
56. Y. Wang and M. Kitsuregawa. Evaluating contents-link coupled web page clustering for web search results. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, pages 499–506, 2002.
57. Y. Wang and M. Kitsuregawa. Enhancing contents-link coupled web page clustering and its evaluation. In *Proceedings of Data Engineering Workshop (DEWS2004)*, 2004.
58. H. White. Citation analysis and discourse analysis revisited. *Applied Linguistics*, 25(1):89–116, 2004.

59. N. Wyse, R. Dubes, and A. Jain. A critical evaluation of intrinsic dimensionality algorithms. *Pattern Recognition in Practice*, pages 415–425, 1980.
60. Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the International Conference on Machine Learning*, pages 412–420, 1997.

APPENDIX A

This appendix provides a detailed explanation of the bi-clustering algorithm used in this paper. So, as a preprocessing step of the bi-clustering algorithm, we have constructed three matrices A, B, C. All of these matrices have documents as row elements, and terms as column elements. For columns, A uses the terms contained in the document collection (the *original terms*), B uses the citation terms, and C uses a combination of citation and original terms. An entry for A(y,x) indicates the number of times term y occurs in document x. An entry for B(y,x) indicates the number of times that document x is cited using term y. An entry for C(y,x) indicates the number of times y is used to describe document x in a citation context or in the original source text for the document x.

Feature Selection. It is well-known that the efficiency of text clustering techniques can suffer from the high dimensionality of the feature space. There are two techniques used to deal with this problem: feature extraction and feature selection. Feature extraction is a process which produces a set of new features from the original features [59]. Feature extraction methods may generate new features which might not have a clear meaning. Therefore, the clustering results are difficult to interpret [12]. Feature selection, on the other hand, is a process that selects only a sub-set from the original feature set based on some criteria. The selected feature keeps the original physical meaning and hence provides a better understanding for the data [31].

In our document clustering, we have tested several unsupervised feature selection methods. Firstly, Document Frequency, DF is the simplest criterion for term selection and easily scales to a large dataset with linear computation complexity [60]. Secondly, Term Contribution, TC is easily biased by those common terms which have high document frequency but uniform distribution over different classes [31]. Thirdly, Mean *TF-IDF*, TI values a term with high term frequency, but low document frequency [52]. Lastly, in Term Frequency Variance (TfV), the "quality" of a term is ranked based on the variance of its term frequency; this is similar in spirit to the intuition of the TI method [14].

In [52], there is a comparison of these unsupervised methods based on preliminary experiments; they state that DF systematically performs worse than TI and TfV, and there is no statistical difference between TI and TfV. In our experiments, we have applied TfV as a feature selection method in order to reduce the dimensionality of our representations (i.e. reduce the number of terms in the previously mentioned matrices A, B and C). This makes it feasible to run our document bi-clustering algorithm.

Bi-clustering Notations. Bi-clustering is a relatively new and less well known approach for clustering. Bi-clustering as already stated is the simultaneous clustering of the rows and columns of a matrix. In this paper, we use it as a technique for identifying sets of documents which all match some set of common terms. More formally:

1. A bi-clustering **BC** consists of k bi-clusters ***bcs***: **BC** = {**bc**1, **bc**2 . . **bc**k}.

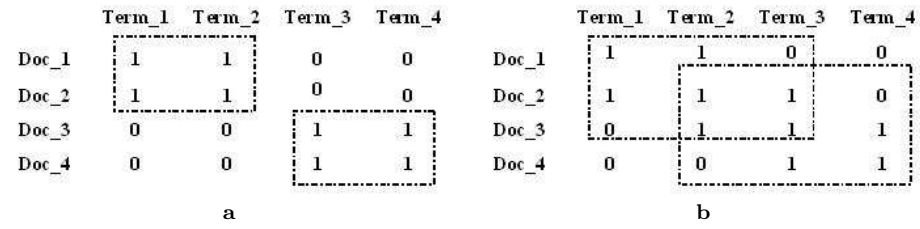

Fig. 8 (a) is an example of Exact Bi-Clustering and (b) is an example of Approximate Bi-Clustering.

2. A bi-cluster $\boldsymbol{bc}$i has a label $\boldsymbol{l}$i of s terms: $\boldsymbol{l}$i = {**t**1, **t**2 . . **t**s}; where 1 < s < m; and m is the length of the whole term vector of the matrix.
3. All bi-cluster label terms {**t**1, **t**2 . . **t**s} must match every document {**d**1, **d**2 . . **d**r} in the bi-cluster[18]; where 1 < r < n; and n is the total number of documents in the dataset.

Bi-clustering based on Exact Frequent Term-sets (EFT). For a given matrix, documents are bi-clustered using a bi-clustering algorithm into a number of groups (bi-clusters). Each group contains all documents matching a particular sub-set of terms [33]. More precisely: the bi-clustering algorithm generates bi-clusters that each contain a subset of the rows of the matrix (the documents) that exhibit the same (exact) behavior across a subset of the columns of the matrix (the terms). Looking at Figure 8(a) which has two bi-clusters, we can see that all documents clustered must have a set of terms in common. For example, Doc_1 and Doc_2 have Term_1 and Term_2 in common. The following is a real example of a bi-cluster (EFT) discovered from our Physics dataset when using the combined representation:

Bi-cluster label terms are "*dimensional, fermion, metric, parameter, function, super-symmetry, theory, vanish, gauge, mass*", which are either contained in or used to refer to 13 documents.

Relaxed Bi-clustering. For text applications such as document clustering, bi-clustering can be too restrictive a process, because it is difficult to find large sets of terms which are relevant to a large group of documents. Consequently, the bi-clusters returned may be small and not sufficiently meaningful to the user. To address this issue, the data mining community has also examined a more relaxed type of biclustering, where the matching constraints do not need to be as strict [30]. This allows the generation of large bi-clusters, containing many documents and many terms in the bi-cluster label. Such large bi-clusters essentially reflect "communities" of documents, whose nature is described by the given terms. The relaxed bi-clustering method used in the experiments described in this paper is based on the use of Approximate Frequent Term-sets (AFT). AFT relaxes the exact matching criterion and allows documents to violate some conditions. For example, AFT allows a document to miss some of the bi-cluster's label terms and thus a term does not have to represent each document in that bi-cluster [30]. In other words, AFT is a relaxed version of EFT. Looking at Figure 8(b) which

[18] By match, we mean that either the document contains the term, or there is a citation to it using this term, depending on which representation is being used

has two bi-clusters, we can see that the relaxation rate is 30% which means that all bi-clusters are allowed to have noise in both rows and columns up to 30%. For example, Doc_1, Doc_2 and Doc_3 have to have at least two terms from the set of terms Term_1, Term_2 and Term_3. Also, the set of terms Term_1, Term_2 and Term_3 have to appear in at least two documents from the set of documents Doc_1, Doc_2 and Doc_3. More formally [30], two conditions required for relaxed bi-clustering are:

- Documents should match most of the terms in the label of the bi-cluster. One allows *Epsilon-r* percent of the terms in the label of the bi-cluster to be missed by a document. Since *Epsilon-r* refers to the percentage of terms which is permitted to be missed, a document is considered for inclusion if it contains (1 - *Epsilon-r*) percent of the total number of the terms in the label of the bi-cluster. In other words, *Epsilon-r* is the permitted noise in each row.
- Terms should match most of the documents in the bi-cluster. One allows *Epsilon-c* percent of documents to be missed. So a term is considered for inclusion if it matches (1 - *Epsilon-c*) percent of the total number of the documents contained in the bi-cluster. In other words, *Epsilon-c* is the permitted noise in each column.

Epsilon-r and *Epsilon-c* are two thresholds that should be specified by the user. In our experiments, we have set both the document and term (rows and columns respectively) thresholds to 25%. . Due to the limited space, we do not describe the bi-clustering algorithm. Here is an example of some term labels generated for one of our bi-clusters in the Physics domain using the AFT approach: *instanton, curve, heterotic, topological, bundle, calabi-yau, intersection, elliptic, threefold, fiber, fold, mirror, scale.*

Summarization and Clustering of Bi-clusters. We have experimentally observed that bi-clustering techniques generate a very large number of bi-clusters. Therefore, we elected to further cluster these bi-clusters using *K-means* clustering, so that a smaller number of clusters would then be obtained and act as a more compact summary. That is, every bi-cluster was represented by a vector consisting of the bi-cluster's label terms and the document IDs it contains. Then a simple K-means clustering was applied in order to reduce the size of the bi-clustering by merging its similar bi-clusters. This helps when we evaluate the accuracy of the bi-clusterings based on the documents' classes. More details about the evaluation process will be discussed later.