Retrieval of Scholarly Articles with Similar Core Contents

Rey-Long Liu*

ARTICLE INFO

Article history:
Received 13 April 2017
Revised 29 July 2017
Accepted 01 August 2017

Keywords:
Scholarly Article,
Article Similarity Estimation,
Core Contents,
Goal Similarity,
Background Similarity,
Conclusion Similarity

ABSTRACT

Retrieval of scholarly articles about a specific research issue is a routine job of researchers to cross-validate the evidence about the issue. Two articles that focus on a research issue should share similar terms in their *core contents*, including their goals, backgrounds, and conclusions. In this paper, we present a technique CCSE (Core Content Similarity Estimation) that, given an article *a*, recommends those articles that share similar core content terms with *a*. CCSE works on titles and abstracts of articles, which are publicly available. It estimates and integrates three kinds of similarity: *goal similarity, background similarity*, and *conclusion similarity*. Empirical evaluation shows that CCSE performs significantly better than several state-of-the-art techniques in recommending those biomedical articles that are judged (by domain experts) to be the ones whose core contents focus on the same research issues. CCSE works for those articles that present research background followed by main results and discussion, and hence it may be used to support the identification of the closely related evidence already published in these articles, even when only titles and abstracts of the articles are available.

1. Introduction

Retrieval of scholarly articles is a fundamental task routinely conducted by researchers. For example, in the biomedical domain, identification of related articles (prioritization of the articles) is essential for evidence collection and validation conducted by biomedical researchers (Lu & Hirschman, 2012). As the retrieval is often motivated by specific research issues, the researchers often strive to find multiple scholarly articles that are closely related to the issues. Therefore, given a scholarly article *a*, several search engines provide the service of recommending those articles that are related to *a* (e.g., Google Scholar at https://scholar.google.com, and PubMed at http://www.ncbi.nlm.nih.gov/pubmed). Many techniques have been developed to estimate the similarity between scholarly articles (e.g., Liu, 2015; Boyack et al., 2011; Aljaber et al., 2010; Gipp & Beel, 2009; Couto et al., 2006). These techniques often worked on titles and abstracts of articles, as well as other text-based information (e.g., main

* Department of Medical Informatics, Tzu Chi University, Hualien, Taiwan (rlliutcu@mail.tcu.edu.tw)
International Journal of Knowledge Content Development & Technology, 7(3): 5-27, 2017.
http://dx.doi.org/10.5865/IJKCT.2017.7.3.005

text of each article) and link-based information, including out-link references (i.e., how an article cites others) and in-link citations (i.e., how an article is cited by others).

In this paper, we extend a preliminary work on estimating inter-article similarity (Liu, 2017). We present and evaluate a novel technique CCSE (Core Content Similarity Estimation) for the inter-article similarity estimation. When compared with the existing search engines and recommendation techniques, CCSE has two interesting features: (1) it works on article titles and abstracts only, which are freely available on the Internet (other parts of the articles, such as the main text, out-link references, and in-link citations, have restricted access); and (2) it improves inter-article estimation by considering the *core contents* of the articles. Core contents of a scholarly article include the research goal, background (problem description), and conclusion of the article. Two articles that share similar terms in their core contents can provide closely related evidence for further research and validation. CCSE can thus be used to improve search engines in recommending those articles with similar core contents, even when only titles and abstracts of the articles are available.

Development of CCSE is challenging, because core contents (research goal, background, and conclusion) of a scholarly article may be briefly expressed in the title and scattered in the abstract. Technical challenges include (1) recognition of the core contents of an article and (2) estimation of the inter-article similarity based on the core contents recognized. We tackle the first challenge by the hypothesis that term t in article a may have different degrees of relatedness to the goal, the background, and the conclusion of a, depending on the *positions* where t appears in the title and the abstract of a. CCSE is developed for those articles that present research background followed by main results and discussion. In these articles, the background mainly introduces the research problem, which tends to appear at the beginning of the abstract; and conversely the conclusion mainly describes the main results, which tend to appear at the end of the abstract. On the other hand, to tackle the second challenge, we employ the hypothesis that mismatch between any parts of the core contents of two articles may significantly reduce the similarity between the two articles, because such mismatch may indicate that the two articles do not focus on the same research issues. Based on the two hypotheses, CCSE separately estimates three kinds of similarity (*goal similarity*, *background similarity*, and *conclusion similarity*), and integrates them to produce the similarity between two articles.

Empirical evaluation is conducted to investigate the contribution of CCSE. It shows that CCSE performs significantly better than several state-of-the-art techniques (including text-based and link-based techniques, as well as their hybrid) in recommending biomedical articles that are judged (by domain experts) to be the ones whose core contents focus on the same research issues. CCSE can thus be used to improve search engines in recommending biomedical articles for further analysis and validation, even when only titles and abstracts of the articles are available.

2. Related Work

To retrieve related articles for a given scholarly article, previous studies have developed many inter-article similarity estimation techniques. These techniques may be *link-based* techniques, which worked on citation relationships among the articles (i.e., out-link references and in-link citations of

each article). They may also be *text-based* techniques, which worked on textual contents of the articles (i.e., title, abstract, keywords, and main text of each article). *Hybrid* techniques were also developed to employ both the text-based and the link-based information. We compare CCSE with these techniques to identify the contribution of CCSE.

2.1 Link-based techniques

To estimate inter-article similarity, previous linked-based techniques employed two kinds of citation links: *in-links* and *out-links*. For an article *a*, in-link citations are those articles that cite *a*, while out-link references are those articles that are cited by *a*. *Co-citation* (Small, 1973) is a representative technique that considers in-links of scholarly articles (Couto et al., 2006). Two articles may be related to each other if they are co-cited by other articles. However, applicability of the techniques based on in-link citations is limited, as many scholarly articles have very few (or even no) in-link citations. CCSE works on titles and abstracts of scholarly articles, which are publicly available.

Another type of link-based techniques are those that worked on out-link references of articles. Out-link references were found to be more helpful than in-link citations in the classification (Couto et al., 2006) and clustering (Boyack & Klavans, 2010) of scholarly articles. *Bibliographic coupling* (BC) is a representative technique that considers out-links (Kessler, 1963). Equation 1 is a typical way to estimate BC similarity between two articles $a1$ and $a2$ (Couto et al., 2006; Calado et al. 2003), where O_{a1} and O_{a2} are the sets of articles that $a1$ and $a2$ cite respectively. When both O_{a1} and O_{a2} are empty, the similarity is set to 0.

$$Similarity_{BC}(a1, a2) = \frac{|O_{a1} \cap O_{a2}|}{|O_{a1} \cup O_{a2}|} \quad (1)$$

BC performed well for classification of scholarly articles (Couto et al., 2006), retrieval of similar legal judgments (Kumar et al., 2011), and detection of plagiarism (Gipp & Meuschke, 2011). However, applicability BC is limited as well, due to two reasons: (1) out-link references of many scholarly articles are not publicly obtainable, and (2) two articles with similar core contents may still cite different articles. We will employ BC as a baseline to show that, by only working on publicly available textual contents (i.e., titles and abstracts) of articles, CCSE can perform significantly better than BC in identifying those articles that share similar core contents.

2.2 Text-based techniques

Text-based techniques worked on textual contents of the articles, which can be titles, abstracts, keywords, and main bodies of the articles. Similarity between two articles are often dominated by those terms (in the two articles) that have higher weights. The weight of a term was often estimated by the frequency and positions of the term in an article, as well as how rarely the term appears in a large collection of articles (a term that appears in fewer articles gets a higher weight). Based on the term weights, many techniques were developed to estimate inter-article similarity. The vector

space model (VSM) is a typical technique. It represents each article as a vector of term weights, and similarity between two articles is simply the cosine similarity on their vectors. VSM was employed by many retrieval systems (e.g., Lucene at http://lucene.apache.org). However, cosine similarity did not perform well in many cases (Whissell & Clarke, 2013; Boyack et al., 2011). Latent Semantic Analysis (LSA) was a typical technique that employed singular value decomposition to improve the vector representation of scholarly articles (Glenisson et al., 2005; Landauer et al., 2004). However LSA did not perform well for scholarly articles either (Boyack et al., 2011). Previous studies thus developed or identified several techniques that had better performance. These techniques included *BM25*, *OK*, and *PubMed*.

BM25 (Robertson et al., 1998) was one of the best techniques in finding related scholarly articles (Boyack et al., 2011). Given an article a_1 as the target, BM25 employs Equation 2 to estimate the score (similarity) of another article a_2 with respect to a_1. In Equation 2, k_1 and b are two parameters, $|a|$ is the number of terms in article a (i.e., length of a), *avgal* is the average number of terms in an article (i.e., average length of articles), $TF(t,a)$ is the frequency of term t appearing in article a, and $IDF(t)$ is the inverse document frequency of term t, which measures how rarely t appears in a large collection of articles

$$Similarity_{BM25}(a_1,a_2)=\sum_{t\in a_1\cap a_2}\frac{TF(t,a_2)(k_1+1)}{TF(t,a_2)+k_1(1-b+b\frac{|a_2|}{avgal})}Log_2IDF(t) \tag{2}$$

OK was developed based on BM25. It was shown to be one of the best techniques to identify related articles as well (Whissell & Clarke, 2013). OK employs Equation 3 to estimate the similarity between two articles a_1 and a_2.

$$Similarity_{OK}(a_1,a_2)=\sum_{t\in a_1\cap a_2}\frac{TF(t,a_1)(k_1+1)}{TF(t,a_1)+k_1(1-b+b\frac{|a_1|}{avgal})}\frac{TF(t,a_2)(k_1+1)}{TF(t,a_2)+k_1(1-b+b\frac{|a_2|}{avgal})}Log_2IDF(t) \tag{3}$$

PubMed was found to be one of the best systems in finding related scholarly articles as well (Boyack et al., 2011). It is a popular search engine for biomedical researchers. Many factors are considered by PubMed to estimate the similarity between two scholarly articles (PubMed, 2014; Lin & Wilbur, 2007), including (1) stemming of the terms in the articles, (2) lengths of the articles (i.e., number of terms in the articles), (3) positions of the terms in the articles (e.g., terms in titles of the articles), (4) key terms of the articles in domain-specific thesauri (e.g., Medical Subject Headings, MeSH, available at http://www.ncbi.nlm.nih.gov/mesh), and (5) weights of the terms in the articles (weighted by the term frequency in an article and how rarely the term appears in a collection of articles).

However, all these state-of-the-art text-based techniques did not consider the similarity between core contents of scholarly articles. Core contents of a scholarly article mainly consist of the research goal, background, and conclusion of the article. They may be briefly expressed in the title and scattered

in the abstract of the article. Our technique CCSE is proposed to recognize the core contents from the title and the abstract of the article, and based on the core contents recognized, estimate the similarity between scholarly articles. We will employ these state-of-the-art text-based techniques (i.e., BM25, OK, and PubMed) as baselines to show that, by estimating inter-article similarity based on the core contents, CCSE can perform significantly better than all the baselines in identifying those articles that share similar core contents.

2.3 Hybrid techniques

Hybrid techniques worked on both link-based and text-based information. They may aim at estimating *inter-article* similarity or *inter-journal* similarity (to estimate inter-journal similarity, citations of articles in a journal were aggregated, Janssens et al., 2009; Liu et al., 2010). We compare CCSE with those techniques that estimated inter-article similarity, as CCSE aims at retrieving articles with similar core contents.

Previous hybrid techniques for inter-article similarity estimation fell into two types: (1) those that relied on *in-link* citations and (2) those that relied on *out-link* references. The first type of hybrid techniques typically considered the *positions* and *context passages* around each out-link citation in *full-text* articles. Positions of a citation in an article may be helpful, because two articles may be similar to each other if they are cited in nearby areas in many articles that cite them (Boyack et al., 2013; Gipp & Beel, 2009). Context passages around a citation x in an article a are the text that authors of a employ to comment x. These passages can thus indicate the main contents of x (Liu et al. 2013), although the citing articles may focus on different parts of x (Elkiss et al., 2008; Kumar et al., 2011) with different sentiments (Small, 2011). The context passages were used for several different purposes, including inter-article similarity estimation (Aljaber et al., 2010), topic-based article retrieval (Liu et al., 2014; Ritchie et al., 2008), and disambiguation of named entities (Nakov et al. 2004). However, applicability of these techniques is limited, due to two reasons: (1) many scholarly articles have very few (or even no) in-link citations, and (2) many scholarly articles do not have publicly obtainable full text. CCSE works on titles and abstracts of scholarly articles, which are publicly available.

Another type of hybrid techniques relied on out-link references. Bibliographic coupling (BC, as noted in Equation 1) was a typical link-based similarity that was integrated with text-based similarity (Liu, 2015; Boyack & Klavans, 2010; Janssens et al., 2008; Couto et al., 2006). The hybrid techniques worked on full-text articles (Liu, 2015; Janssens et al., 2008) or only titles and abstracts of the articles (Boyack & Klavans, 2010; Couto et al., 2006). As full-text articles are often not publicly obtainable, we compare CCSE with those techniques that only worked on titles and abstracts of articles. These hybrid techniques did not necessarily perform significantly better than BC (Couto et al., 2006). One of the hybrid techniques performed better than BC in some cases (Boyack & Klavans, 2010). It treated a co-reference cited by two articles as a co-word in the two articles. We will thus implement the hybrid technique as a baseline to investigate the contribution of CCSE.

Therefore, when compared with the link-based, text-based, and hybrid techniques for retrieving similar scholarly articles, CCSE has two contributions: (1) it works on publicly obtainable parts of

the articles (i.e., titles and abstracts of the articles), making it able to recommend similar articles more comprehensively, and (2) it estimates the similarity between two scholarly articles based on their core contents, making it able to recommend those articles that share similar core contents. We will show that CCSE can perform significantly better than several baselines in identifying those biomedical articles whose core contents focus on the same research issues.

3. Core Content Similarity Estimation

Given two articles a_1 and a_2, CCSE estimates the similarity between them based on their titles and abstracts. Technical challenges of CCSE include (1) recognition of the core contents of a_1 and a_2, and (2) estimation of the similarity based on the core contents recognized. To tackle the first challenge, CCSE employs the hypothesis that a term t in an article a may have different degrees of relatedness to the *background*, the *conclusion*, and the *goal* of a, depending on the *positions* where t appears in the title and the abstract of a. Figure 1 shows two linear ways to respectively estimate the relatedness of a term to the background (R_{back}) and the conclusion (R_{conc}) of an article a. The research background mainly defines the research problem (of a), which tends to appear at the beginning of the abstract of a; and conversely the conclusion mainly describes the main findings (of a), which tend to appear at the end of the abstract of a. Similarly, Figure 2 shows a way to estimate the relatedness of a term to the research goal of a (R_{goal}), which tends to appear at the title of a, as well as the beginning and the end of the abstract of a, because the research goal may describe both the research problem and the main findings of a. The idea is employed to estimate how a term in an article is related to the core contents of the article.

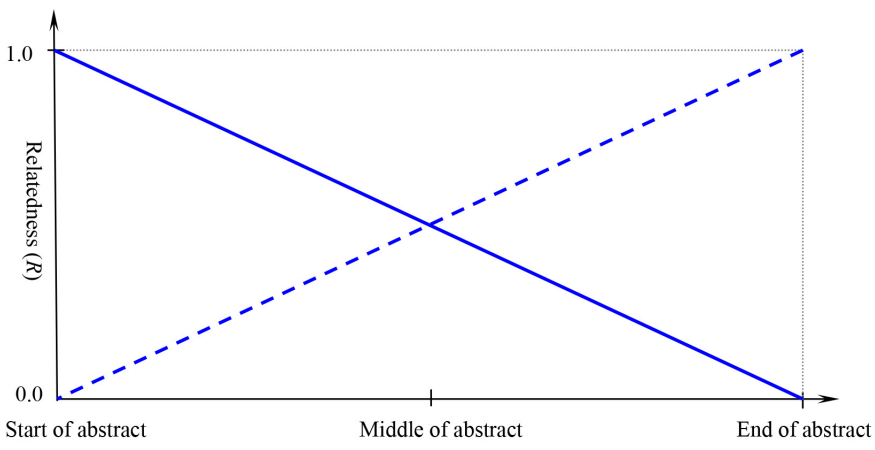


Fig. 1. Relatedness of a term to the *background* (R_{back}, the solid line) and the *conclusion* (R_{conc}, the dashed line) of a scholarly article: The background mainly defines the research problem, which tends to appear at the beginning of the abstract of the article; and conversely the conclusion mainly describes the main results, which tend to appear at the end of the abstract

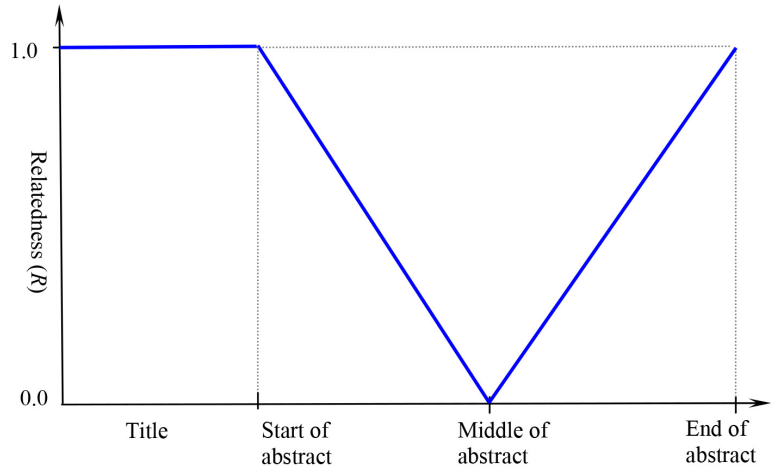


Fig. 2. Relatedness of a term to the *goal* (R_{goal}) of a scholarly article: The goal tends to express the research problem and results, which tend to appear at the title of the article, as well as the beginning and the end of the abstract of the article

Moreover, to tackle the second challenge (estimation of the similarity based on the core contents recognized), CCSE employs the hypothesis that mismatch between any parts of the core contents of two articles a_1 and a_2 may significantly reduce the similarity between a_1 and a_2, because such mismatch may indicate that a_1 and a_2 do not focus on the same research issues. Table 1 illustrates the main idea. A term t may have a *strong* effect on the similarity between a_1 and a_2 in several cases (see *Type I effect* in Table 1): (1) it may significantly *increase* the similarity between a_1 and a_2 if it appears in both a_1 and a_2 and is related to the core contents of a_1 and a_2, and (2) it may significantly *decrease* the similarity between a_1 and a_2 if it is related to the core content of a_1 (a_2) but does not appear in a_2 (a_1). Moreover, a term t may have a *little* effect on the similarity between a_1 and a_2 (see *Type II effect* in Table 1) if it appears in both a_1 and a_2 and is related to the core content of a_1 (a_2) but *not* related to the core content of a_2 (a_1). Finally, effects of certain terms may be ignored (see *Type III effect* in Table 1), including those that are *not* related to core contents of a_1 and a_2; those that do not appear in a_1 and a_2; as well as those that are *not* related to the core content of a_1 (a_2) and do not appear in and a_2 (a_1).

CCSE is developed based on the above two hypotheses. Similarity between two articles a_1 and a_2 is estimated by Equation 4, which separately checks how the core content of a_1 appears in a_2 (i.e., $CoreMatch(a_1, a_2)$) and vice versa (i.e., $CoreMatch(a_2, a_1)$). CCSE thus conducts a *dual* match so that a_1 and a_2 are said to be quite similar to each other only if the core content of a_1 appears in a_2 *and* the core content of a_2 appears in a_1 as well. Any mismatch between the core contents will significantly reduce the similarity between a_1 and a_2.

$$Similarity_{CCSE}(a_1, a_2) = CoreMatch(a_1, a_2) \times CoreMatch(a_2, a_1) \quad (4)$$

Table 1. A term t may have three types of effects (*strong effect*, *little effect*, and *no effect*) on the similarity between two articles a_1 and a_2, depending on how t is related to the core contents of a_1 and a_2

			(for article a_2)		
			t appears in a_2		t does *not* appear in a_2
			Core	*Non-core*	
(for article a_1)	t appears in a_1	*Core*	**Type I:** *strong effect*	**Type II:** *little effect*	**Type I:** *strong effect*
		Non-core	**Type II:** *little effect*	**Type III:** *no effect*	**Type III:** *no effect*
	t does *not* appear in a_1		**Type I:** *strong effect*	**Type III:** *no effect*	**Type III:** *no effect*

To estimate how the core content of a_1 appears in a_2, CCSE employs Equation 5, which separately estimates how the three parts of the core content of a_1 (*goal*, *background*, and *conclusion* of a_1) appear in a_2. The degrees of match in the three parts are averaged to produce the degree of match of the core content.

$$CoreMatch(a_1, a_2) = \frac{Match_{goal}(a_1, a_2) + Match_{back}(a_1, a_2) + Match_{conc}(a_1, a_2)}{3} \tag{5}$$

Figure 3 illustrates the way to estimate how the three core parts of a_1 match those of a_2. CCSE estimates how core content terms of a_1 appears in a_2 as core content terms. Equation 6 is defined to estimate how the *goal* of a_1 appears in a_2. As the title of a_1 can provide the most reliable information about the goal of a_1, Equation 6 checks how the terms in the title of a_1 appear in a_2. When a term t in the title of a_1 appears in a_2, it increases the degree of goal match between a_1 in a_2 (see the numerator of Equation 6). The increment of the similarity is based on the degrees of relatedness of t to the goals of a_1 and a_2 (i.e., $R_{goal}(t,a_1)$ and $R_{goal}(t,a_2)$, with the smaller one as the similarity increment, ref. Equation 7). These degrees of relatedness are defined based on the hypothesis discussed above (see Figure 2). If t appears at multiple positons in an article, its degree of relatedness is set to the maximum degree of relatedness at these positions.

$$Match_{goal}(a_1, a_2) = \frac{\sum_{t \in Title(a_1); t \in Title(a_2) \cup Abstract(a_2)} InterR(R_{goal}(t, a_1), R_{goal}(t, a_2)) \times Log_2 IDF(t)}{\sum_{t \in Title(a_1)} R_{goal}(t, a_1) \times Log_2 IDF(t)} \tag{6}$$

$$InterR(r_1, r_2) = \begin{cases} r_1, & if\ r_1 < r_2; \\ r_2, & otherwise. \end{cases} \tag{7}$$

$$IDF(t) = \frac{1 + Total\ number\ of\ articles}{1 + Number\ of\ articles\ containing\ t} \tag{8}$$

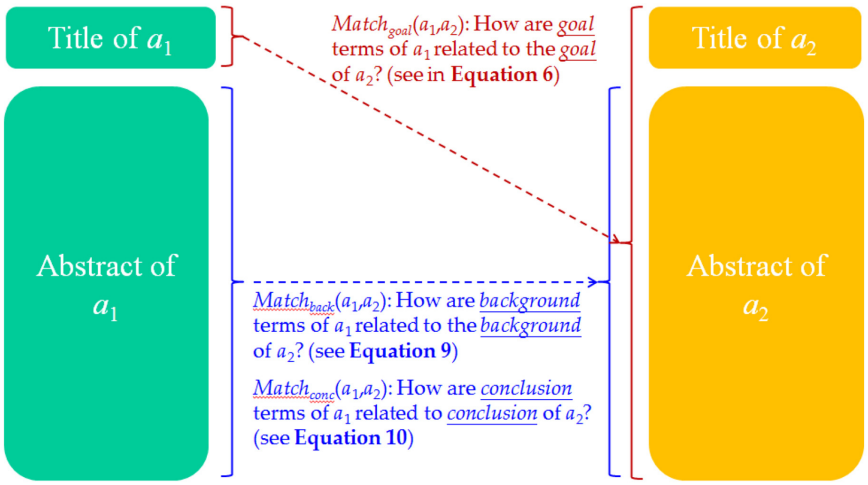


Fig. 3. Estimation of how the three core parts (*goal*, *background*, and *conclusion*) of an article a_1 match those of another article a_2: The estimation is based on how core content terms of a_1 appears in a_2 as core content terms. The degrees of match in the three parts are defined in Equations 6, 9, and 10, respectively

Similarly, Equation 9 is defined to estimate how the *background* of a_1 appears in a_2 (i.e., $Match_{back}$), and Equation 10 is defined to estimate how the *conclusion* of a_1 appears in a_2 (i.e., $Match_{conc}$). The hypothesis discussed above (see Figure 1) is employed to estimate the degrees of relatedness of a term t to the background of a_1 and a_2 (i.e., $R_{back}(t,a_1)$ and $R_{back}(t,a_2)$) as well as the conclusion of a_1 and a_2 (i.e., $R_{conc}(t,a_1)$ and $R_{conc}(t,a_2)$). If the abstract of a_1 or a_2 is not available, $Match_{back}$ and $Match_{conc}$ are set to $Match_{goal}$ so that all articles may be ranked for researchers to check, even when the abstract is not publicly available.

$$Match_{back}(a_1,a_2)=\begin{cases} Match_{goal}(a_1,a_2), \ if\ a_1\ or\ a_2\ has\ no\ abstract \\ \dfrac{\sum_{t\in Abstract(a_1);t\in Abstract(a_2)} InterR(R_{back}(t,a_1),R_{back}(t,a_2))\times Log_2 IDF(t)}{\sum_{t\in Abstract(a_1)} R_{back}(t,a_1)\times Log_2 IDF(t)}, otherwise. \end{cases} \quad (9)$$

$$Match_{conc}(a_1,a_2)=\begin{cases} Match_{goal}(a_1,a_2), \ if\ a_1\ or\ a_2\ has\ no\ abstract \\ \dfrac{\sum_{t\in Abstract(a_1);t\in Abstract(a_2)} InterR(R_{conc}(t,a_1),R_{conc}(t,a_2))\times Log_2 IDF(t)}{\sum_{t\in Abstract(a_1)} R_{conc}(t,a_1)\times Log_2 IDF(t)}, otherwise. \end{cases} \quad (10)$$

Therefore, given the title and the abstract of a scholarly article, CCSE considers three parts of the core content of the article: goal, background, and conclusion of the article. CCSE estimates how each term is related to each of the three parts based on the positions of the term appearing in the

title and the abstract of the article. A dual match is then conducted to estimate the similarity between two articles so that any mismatch between the core contents of the two articles will significantly reduce the similarity between them. CCSE is thus a pure text-based technique that works on titles and abstracts of scholarly articles, which are publicly available on the Internet.

4. Experiments

CCSE is evaluated in two experiments on real-world data. We measure the contribution of CCSE in identifying those scholarly articles whose core contents focus on the same research issues.

4.1 The data

The experimental data was collected from DisGeNET (at http://www.disgenet.org/web/DisGeNET/menu/home), which maintains a database of gene-disease associations. We test 53 gene-disease pairs that had the largest number of related articles annotated by GAD (Genetic Association Database, available at http://geneticassociationdb.nih.gov) or CTD (Comparative Toxicogenomics Database, available at http://ctdbase.org) for human. Both GAD and CTD recruit domain experts to manually select articles to annotate each gene-disease pair (Wiegers et al., 2009; Becker et al., 2004). The articles used to annotate a gene-disease pair <*g*, *d*> thus share similar core contents about the same research issue (i.e., association between *g* and *d*). For each gene-disease pair <*g*, *d*>, we designate one article as the *target*, while the others as the *candidates*. Given the target article, a better technique should rank high these candidates, among other candidate articles that are *not* dedicated to <*g*, *d*>.

For each gene-disease pair <*g*, *d*>, we thus also need to collect many candidate articles that are *not* dedicated to <*g*, *d*>. These candidate articles are "near-miss" articles for <*g*, *d*>, as they were collected by sending two queries to PubMed Central (at http://www.ncbi.nlm.nih.gov/pmc): "*g* NOT *d*" and "*d* NOT *g*". The articles collected by this way mention *g* or *d* but not both, and hence they should not focus on the research issue about the association between *g* and *d*. For each gene-disease pair, at most 200 near-miss candidate articles were collected. A gene-disease pair corresponds to a test in the experiment (we thus have 53 tests in the experiment). There are 9,875 candidate articles among which 135 articles share similar core contents with their respective target articles. The average percentage of the articles that share similar core contents is 1.34%. It is thus challenging to rank these articles high, among the near-miss candidate articles.

4.2 The baselines

Several state-of-the-art techniques are employed as the baselines to investigate the contribution of CCSE. As noted in *Related Work*, previous techniques that estimated the similarity between two scholarly articles a_1 and a_2 can be *link-based* (those that worked on citation relationships among articles), *text-based* (those that worked on textual contents of a_1 and a_2), or *hybrid* (those that worked on both text-based and link-based information). We design two experiments (Experiment I and Experiment

II) in which several text-based, link-based, and hybrid techniques are tested.

In Experiment I, the baselines include a link-based technique (bibliographic coupling, BC), two text-based techniques (BM25 and OK), and a hybrid technique (HybridK50). Note that, in practice, those baselines that rely on citation links (i.e., the link-based and hybrid baselines) may not work when the link-based information (bibliography in articles) is not publicly obtainable. These baselines are employed as baselines simply because we aim at showing that, by relying only on publicly available contents (i.e., titles and abstracts of articles), CCSE can perform even better in identifying those scholarly articles that share similar core contents.

The link-based baseline BC (ref. Equation 1) relies on out-link references. As noted in *Related Work*, BC is employed as a baseline based on three reasons: (1) most articles have out-link references, (2) out-link references were found to be more helpful than in-link citations in the clustering and classification of scholarly articles, and (3) BC was found to be good in retrieving similar documents.

The text-based baselines are BM25 and OK. As noted in *Related Work*, BM25 and OK were two best techniques to find related articles. BM25 estimates the similarity between two articles with Equation 2, with the two parameter k_1 and b being typically set to 2 and 0.75 respectively (Boyack et al., 2011, Liu and Huang, 2011). OK employs Equation 3 to estimate the similarity between two articles, with the two parameters k_1 and b being set to 8 and 1.0 respectively (as suggested by Whissell & Clarke, 2013). As CCSE, both BM25 and OK work on titles and abstracts of the articles.

The hybrid baseline is HybridK50, which performed better than BC in certain cases (Boyack & Klavans, 2010). Similarity between two articles a_1 and a_2 is defined based on the intersection of words and out-link references in a_1 and a_2. HybridK50 estimates the similarity by treating a reference co-cited by a_1 and a_2 as a co-word in the titles and abstracts of a_1 and a_2 (for detailed definition of the similarity measure, the reader is referred to Boyack & Klavans, 2010).

In Experiment II, we compare CCSE with PubMed, which provides the service of "Related Citations" that recommends related articles. As noted in *Related Work*, by employing domain-specific thesauri and integrates several factors about terms, PubMed was found to be one of the best in finding related scholarly articles. We aim at showing that CCSE can perform better than PubMed by focusing on the recognition of core contents of scholarly articles. For each target article a, PubMed recommends a sequence S of related articles. We remove from S all articles that are *not* candidate articles. As noted above (see *The data*), the candidate articles include those that share similar core contents with a (judged by domain experts), as well as those that should *not* share similar core contents with a. By focusing on these candidate articles in S, performance of CCSE and PubMed can be objectively compared.

4.3 Evaluation criteria

Two evaluation criteria are used to measure the performance of CCSE and the baselines. They are *Mean average precision* (MAP) and average P@X, which were routinely employed in text ranking studies. MAP is defined in Equation 11, where $|T|$ is the number of tests (recall that we have 53 tests), and $AvgP(i)$ is the average precision in the i^{th} test. MAP is simply the average of the $AvgP$ values in all the tests.

$$MAP = \frac{\sum_{i=1}^{|T|} AvgP(i)}{|T|} \tag{11}$$

$$AvgP(i) = \frac{\sum_{j=1}^{h_i} \frac{j}{Seen_i(j)}}{h_i} \tag{12}$$

$AvgP(i)$ is defined in Equation 12, where h_i is the number of articles that are judged (by domain experts) to be the ones that focus on the same research issue as the target article in the i^{th} test, and $Seen_i(j)$ is the number of articles that readers have seen when the j^{th} core-content-sharing article in the i^{th} test is shown (i.e., number of articles whose ranks are higher than or equal to that of the j^{th} core-content-sharing article in the i^{th} test). Therefore, given a target article r in the i^{th} test, if a system can rank higher those articles that share core contents with r, $AvgP(i)$ will be higher.

Instead of working on *all* articles (as MAP does), average P@X only works on those articles that are ranked at top-X positions. Average P@X is defined in Equation 13. It is the average of the P@X values in all the 53 tests. Equation 14 defines P@X, which is the precision when top-X articles are shown to the readers. Therefore, when X is set to a small value, P@X measures how a system ranks core-content-sharing articles very high. In the experiments, we set X to 1, 3, and 5.

$$\text{Average P@X} = \frac{\sum_{i=1}^{|T|} \text{P@X}(i)}{|T|} \tag{13}$$

$$\text{P@X}(i) = \frac{\text{Number of top-X articles that share core contents with the target in the } i^{th} \text{ test}}{\text{X}} \tag{14}$$

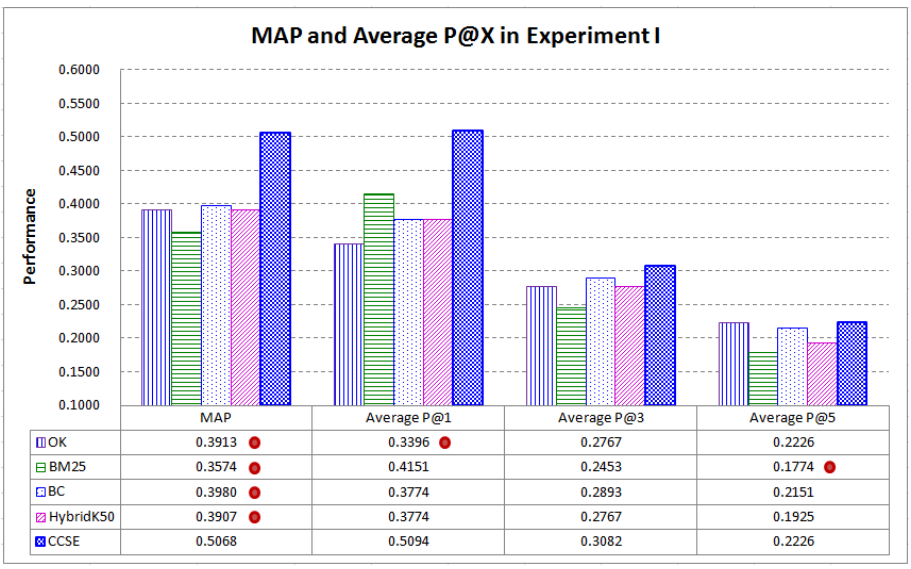
Moreover, to verify whether the performance differences between CCSE and each of the baselines are *statistically significant*, we conduct significance tests by two-sided and paired t-tests with 95% confidence level. The significance tests are conducted on the *AvgP* and P@X values in all the 53 tests.

4.4 Results

Figure 4 compares performance of CCSE and the four state-of-the-art baselines (OK, BM25, BC, and HybridK50). CCSE performs better than all the baselines, especially in MAP and average P@1. CCSE is thus more capable of ranking core-content-sharing articles at top-1. MAP of CCSE

is significantly better than MAP of each baseline. The baseline BC, which employs out-link references to rank scholarly articles (ref. Equation 1), achieves the best MAP among the baselines, however CCSE performs 27% better than BC in MAP (0.5068 vs. 0.3980). CCSE performs significantly than BC even though CCSE only works on titles and abstracts of articles, which are more publicly obtainable than bibliography (i.e., out-link references in the articles). Moreover, the baseline BM25, which employs text-based information to rank scholarly articles (ref. Equation 2), achieves the best average P@1 among the baselines, however CCSE performs 23% better than BM25 in average P@1 (0.5094 vs. 0.4151). These contributions are of practical significance to the retrieval of scholarly articles that focus on similar research issues.

It is interesting to note that the link-based baseline BC, the text-based baseline OK, and the hybrid baseline HybridK50 achieve similar performance in MAP, although they have somewhat different performance in average P@X. Therefore, link-based information and text-based information contribute almost equally in the retrieval of core-content-sharing articles, and integration of them is *not* necessarily helpful. This result indicates that improvement of these baselines is not a trivial task. The significantly better performance of CCSE demonstrates that the improvement task can be realized by focusing on the recognition of core contents of scholarly articles.

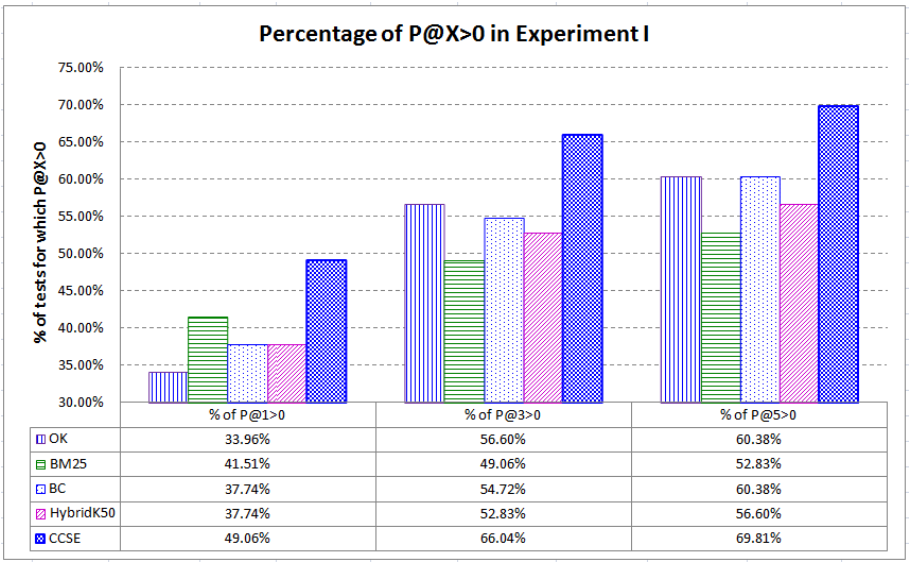


	MAP	Average P@1	Average P@3	Average P@5
OK	0.3913 ●	0.3396 ●	0.2767	0.2226
BM25	0.3574 ●	0.4151	0.2453	0.1774 ●
BC	0.3980 ●	0.3774	0.2893	0.2151
HybridK50	0.3907 ●	0.3774	0.2767	0.1925
CCSE	0.5068	0.5094	0.3082	0.2226

Fig. 4. MAP and average P@X in Experiment I: CCSE performs significantly better than all the baselines in MAP (a dot on a system indicates that performance difference between the system and CCSE is statistically significant)

Figure 5 shows the *percentage* of the tests in which P@X>0 achieved by CCSE and the baselines. A higher percentage achieved by a system indicates that the system is capable of ranking core-content-sharing articles at top positions in more of the 53 tests. Such a system should be preferred in practice, as it demonstrates both good and stable performance in recommending core-content-sharing

articles for different research issues. Among the baselines, BM25 tends to achieve higher percentage in P@1, while OK and BC tend to achieve higher percentage in P@3 and P@5. CCSE achieves the highest percentage. When compared with the best baselines, CCSE contributes 18% improvement in the percentage in P@1 (49.06% vs. 41.51%), 17% improvement in the percentage in P@3 (66.04% vs. 56.60%), and 16% improvement in the percentage in P@5 (69.81% vs. 60.38%). The contribution is of practical significance to researchers, who often care about different research issues, and for each research issue, need to check a large number of scholarly articles.



	% of P@1>0	% of P@3>0	% of P@5>0
OK	33.96%	56.60%	60.38%
BM25	41.51%	49.06%	52.83%
BC	37.74%	54.72%	60.38%
HybridK50	37.74%	52.83%	56.60%
CCSE	49.06%	66.04%	69.81%

Fig. 5. Percentage of the tests in which P@X > 0 in Experiment I: CCSE ranks the core-content-sharing articles at top-1, top-3, and top-5 for a higher percentage of tests than all the baselines

Figure 6 compares performance of CCSE and PubMed in MAP and average P@X. CCSE performs better than PubMed in all evaluation criteria, with statistically significant improvement in average P@3 (7% improvement, 0.5472 vs. 0.5094). Moreover, Figure 7 shows the *percentage* of the tests in which P@X>0 achieved by CCSE and PubMed. CCSE contributes 10% improvement in the percentage in P@1 (83.02% vs. 75.47%), 6% improvement in the percentage in P@3 (94.34% vs. 88.68%), and 2% improvement in the percentage in P@5 (100% vs. 98.11%). These contributions are of practical significance as well, as PubMed is a popular search engine. They are also of technical significance, as PubMed has employed domain-specific thesauri and several typical text-based factors about terms and articles. CCSE achieves better performance by core content recognition. It can be applied to various domains, as it achieves the better performance without relying on any domain-specific thesauri.

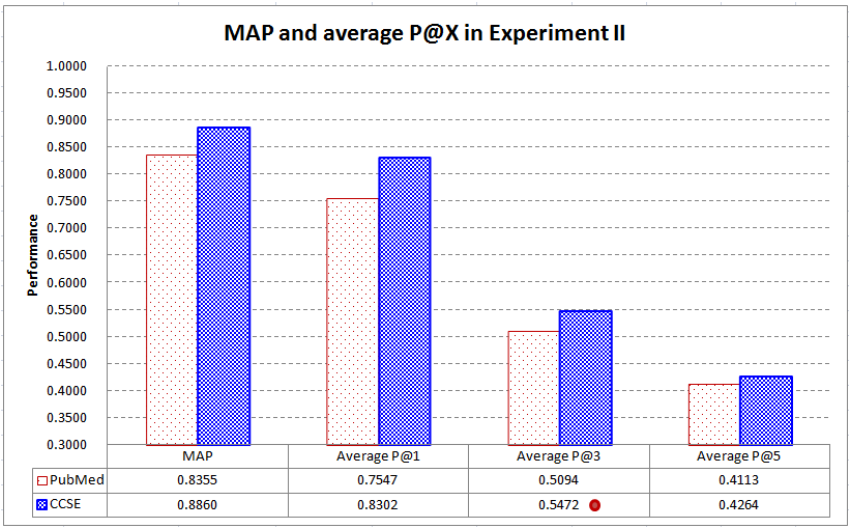


Fig. 6. MAP and average P@X in Experiment II: CCSE performs better than PubMed in all evaluation criteria (a dot on a system indicates that performance difference between PubMed and CCSE is statistically significant)

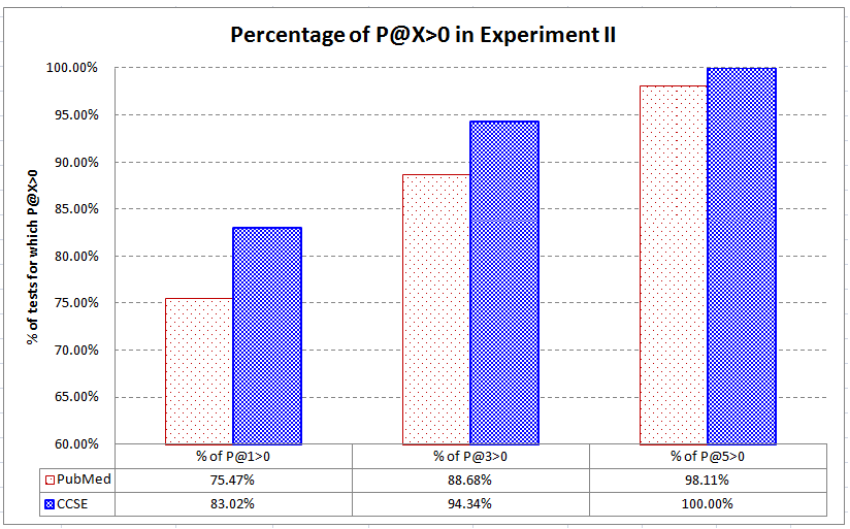


Fig. 7. Percentage of the tests in which P@X > 0 in Experiment II: CCSE ranks core-content-sharing articles at top-1, top-3, and top-5 for a higher percentage of tests than PubMed

With the above experimental results, we summarize several main findings as follows:

(1) CCSE achieves significantly better performance than all the baselines, even though it only works on titles and abstracts of articles, which are more publicly obtainable than bibliography that is employed by the link-based and hybrid baselines. Therefore, even when only the titles and the abstracts are available, retrieval of core-content-sharing

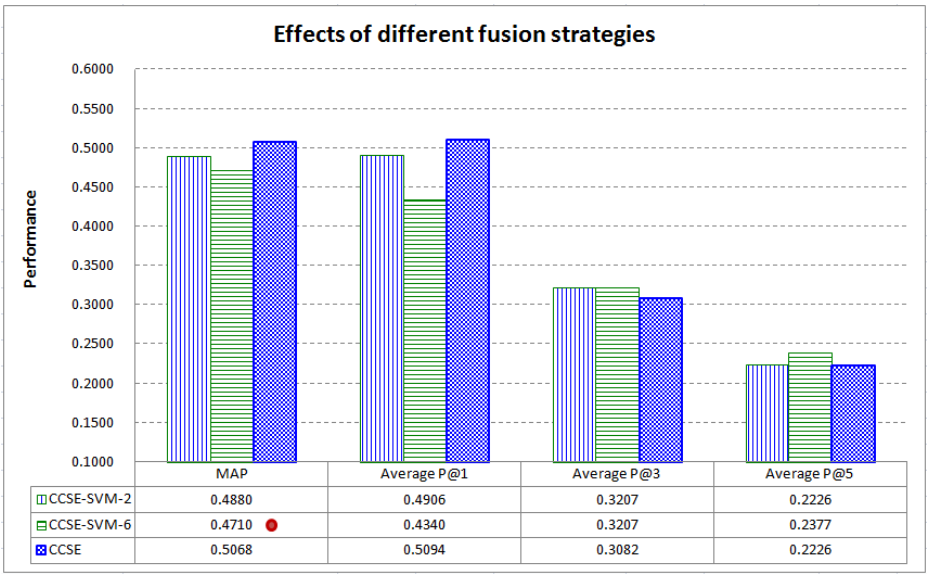
articles can still be significantly improved.

(2) The text-based baseline OK, the link-based baseline BC, and the hybrid baseline HybridK50 achieve similar MAP. A previous study also showed that HybridK50 had similar performance as BC, although it performed better than BC in certain cases (Boyack & Klavans, 2010). Therefore, significant improvement of these baselines is not a trivial task. The significantly better performance of CCSE thus demonstrates that the improvement task can be realized by focusing on the recognition of core contents of scholarly articles.

(3) CCSE performs better than PubMed, which has employed domain-specific thesauri and considered several typical text-based factors. Recognition of core contents of scholarly articles (as CCSE does) is thus more helpful in identifying those articles that focus on similar research issues. CCSE does not rely on any domain-specific thesauri. It can be applied to those articles that present research background followed by main results and discussion.

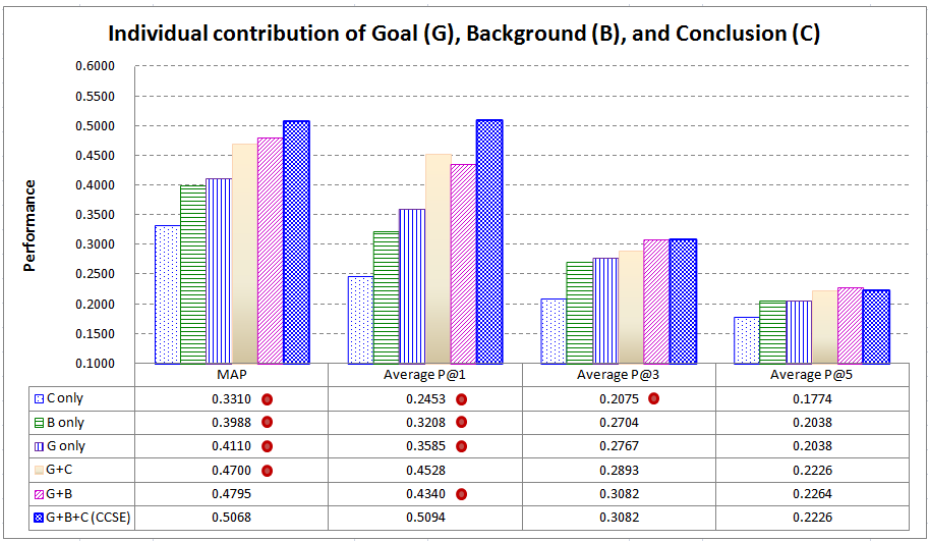
Performance of CCSE is further analyzed from three perspectives. The first perspective is concerned with the possible contribution of employing machine-learning techniques to integrate the similarity factors of CCSE (rather than using Equation 4 and Equation 5). We thus employ SVM (Support Vector Machine) for ranking (Joachims, 2002) as the machine learning technique to integrate the similarity factors. SVM was one of the best techniques routinely used to integrate multiple factors to achieve better ranking (e.g., Liu & Huang, 2011; Veloso et al., 2008). To set up the data for SVM, we conduct 5-fold experiments in which the 53 tests are evenly divided into five parts, and the data in each part serves as the test data for exactly once with remaining data serving as the training data, and the process repeats five times. We try two different fusion strategies: (1) CCSE-SVM-2: two factors $CoreMatch(a_1,a_2)$ and $CoreMatch(a_2,a_1)$ defined in Equation 5 are integrated by SVM; and (2) CCSE-SVM-6: all six factors are integrated by SVM, including $Match_{goal}(a_1,a_2)$ and $Match_{goal}(a_2,a_1)$ defined in Equation 6, $Match_{back}(a_1,a_2)$ and $Match_{back}(a_2,a_1)$ defined in Equation 9, and $Match_{conc}(a_1,a_2)$ and $Match_{conc}(a_2,a_1)$ defined in Equation 10. Experimental results in Figure 8 show that machine-learning-based factor fusion by SVM does not perform better than CCSE. CCSE-SVM-6 even performs significantly worse than CCSE in MAP. The fusion strategy of CCSE (Equation 4 and Equation 5) is thus appropriate, and it may not be necessary to integrate the similarity factors by machine learning.

The second perspective is concerned with the individual contribution of each of the three kinds of similarity considered by CCSE, including goal similarity ($Match_{goal}$ defined in Equation 6), background similarity ($Match_{back}$ defined in Equation 9), and conclusion similarity ($Match_{conc}$ defined in Equation 10). We thus implement three versions '*G only*', '*B only*', and '*C only*' for which *CoreMatch* defined in Equation 5 is modified to respectively consider goal similarity, background similarity, and conclusion similarity only. Experimental results in Figure 9 show that goal similarity tends to be more helpful than the other two kinds of similarity (i.e., '*G only*' performs better than '*B only*' and '*C only*'). We thus implement two additional versions '*G+C*' and '*G+B*'. For '*G+C*', *CoreMatch* defined in Equation 5 is modified to be the average of goal similarity and conclusion similarity. For '*G+B*',



	MAP	Average P@1	Average P@3	Average P@5
CCSE-SVM-2	0.4880	0.4906	0.3207	0.2226
CCSE-SVM-6	0.4710 ●	0.4340	0.3207	0.2377
CCSE	0.5068	0.5094	0.3082	0.2226

Fig. 8. Effects of different strategies to integrate the similarity factors: Machine-learning-based integration by SVM does not perform better than CCSE, and it may even perform significantly worse than CCSE (a dot on a system indicates that performance difference between the system and CCSE is statistically significant)

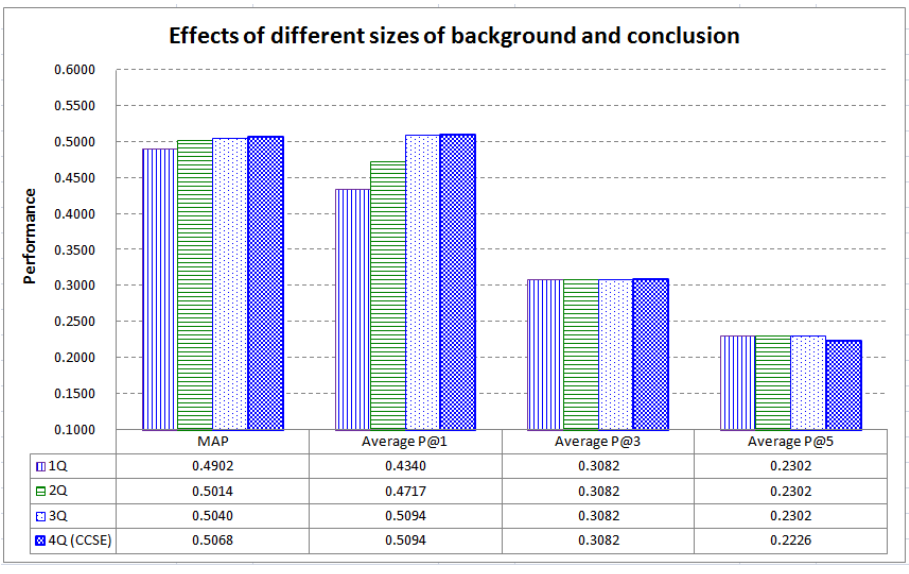


	MAP	Average P@1	Average P@3	Average P@5
C only	0.3310 ●	0.2453 ●	0.2075 ●	0.1774
B only	0.3988 ●	0.3208 ●	0.2704	0.2038
G only	0.4110 ●	0.3585 ●	0.2767	0.2038
G+C	0.4700 ●	0.4528	0.2893	0.2226
G+B	0.4795	0.4340 ●	0.3082	0.2264
G+B+C (CCSE)	0.5068	0.5094	0.3082	0.2226

Fig. 9. Individual contribution of goal similarity, background similarity, and conclusion similarity: Goal similarity tends to be more helpful than the other two kinds of similarity; and simultaneously employing all the three kinds of similarity (as CCSE does) achieves the best performance, especially in MAP and average P@1 (a dot on a system indicates that performance difference between the system and CCSE is statistically significant)

CoreMatch is modified to be the average of goal similarity and background similarity. The results shown in Figure 9 indicate that it is helpful to integrate goal similarity with conclusion similarity or background similarity (i.e., '*G+C*' and '*G+B*' perform better than '*G only*', '*B only*', and '*C only*'). However, CCSE (i.e., '*G+B+C*' in Figure 9) performs significantly better than all of them, especially in MAP and average P@1. Simultaneously employing all the three kinds of similarity is thus a good way to achieve the best performance.

The third perspective is concerned with the effects of setting different sizes for the background area and the conclusion area in an article. As noted above, CCSE employs two linear ways (defined in Figure 1) to estimate the relatedness of a term to the background (R_{back}) and the conclusion (R_{conc}) of an article. Although these two ways give terms different degrees of relatedness based on their positions in the article, they consider the whole article as the potential area in which the terms may appear. We thus implement three versions that restrict the size of the areas: '*1Q*', '*2Q*', and '*3Q*', which consider 1/4, 2/4, and 3/4 of the article only (e.g., in '*1Q*', the background terms can only appear at the first quarter of the article and the conclusion terms can only appear at the last quarter of the article). Experimental results in Figure 10 show that setting a larger area tends to be helpful. Setting the whole article as the potential area (i.e., '*4Q*' as done by CCSE) can achieve the best performance, although the performance differences are not statistically significant. The results justify our expectation that terms about the background and the conclusion of a scholarly article may scatter in the abstract (as noted in *Introduction*). It is thus quite difficult to exactly extract these terms from the abstract, and hence CCSE considers all terms in the abstract and estimates their degrees of relatedness to the core content of the article.



	MAP	Average P@1	Average P@3	Average P@5
1Q	0.4902	0.4340	0.3082	0.2302
2Q	0.5014	0.4717	0.3082	0.2302
3Q	0.5040	0.5094	0.3082	0.2302
4Q (CCSE)	0.5068	0.5094	0.3082	0.2226

Fig. 10. Effects of different sizes of the background and conclusion areas: Considering a larger size tends to achieve better performance, although the performance differences are not statistically significant

5. Discussion

5.1 Application and suggestion

We have shown that CCSE performs well in recommending core-content-sharing biomedical articles, even when only titles and abstracts of the articles are available. We expect that CCSE may be applied to other domains, as it does not relying on any domain-specific thesauri. Given those articles that present research background followed by main results and discussion, CCSE may be invoked to support various kinds of research tasks, including the retrieval, clustering, mining, validation, and curation of the closely related evidence already published in these articles.

To implement the applications, we suggest that the ideas of CCSE should be implemented in search engines of scholarly articles (e.g., Google Scholar and PubMed). These search engines have routinely collected a huge amount of scholarly articles online. The articles are often preprocessed for subsequent retrieval. Typical preprocessing tasks include term indexing and article relatedness estimation. These tasks are essential for the efficient retrieval of related articles online. We suggest that the term relatedness weighting technique of CCSE (ref. Figure 1 and Figure 2) should be incorporated to the term indexing module of the search engines, and the similarity estimation technique of CCSE (ref. Equation 4 to Equation 10) should be incorporated to the article relatedness estimation module of the search engines. With the ideas of CCSE, the search engines can be more capable of recommending core-content sharing articles for a given scholarly article.

5.2 Future work

It is interesting to investigate the performance of CCSE on other datasets and domains. It is also interesting to further investigate the possible contribution of employing different ways to estimate the relatedness of terms to the core content of a scholarly article. CCSE currently employs linear weighting to estimate the relatedness (ref. Figure 1 and Figure 2). One alternative is to employ more complicated language processing techniques or templates to recognize the parts respectively dedicated to the background, goal, and conclusion of an article. Cost-effectiveness of this alternative needs to be investigated in more experiments.

Another interesting future work is to investigate the potential contribution of incorporating link-based information to CCSE. CCSE currently works on titles and abstracts of articles only. Although this way improves the applicability of CCSE (as the titles and abstracts are publicly available), it might still be helpful to consider link-based information (e.g., bibliography) of certain articles for which the information is publicly obtainable (e.g., link-based information of open-access articles). It is thus interesting to explore several issues, including (1) identification of the link-based information that may indicate the core content of an article, (2) automatic recognition of the link-based information in the article, (3) inter-article similarity estimation based on the link-based information, (4) intelligent integration of the link-based similarity with the text-based similarity currently estimated by CCSE, and (5) appropriate ranking of scholarly articles that may or may not have the link-based similarity.

The ideas of CCSE can also be extended to extract *core entity terms* in a scholarly article. Core entity terms in an article a are those terms (in a) that are related to the core content of a. By extracting the core entity terms, several interesting applications can be supported, including the retrieval of those articles that focus on a given entity, as well as the navigation on a network of interacting entities already published in literature, with specific scholarly articles annotated for reader to check. To extract core entity terms from an article, two research issues should be addressed, including the definition of the core entity terms (e.g., how should they be mentioned in the goal, background, and conclusion of the article), and the automatic recognition of the core entity terms. Proper extraction of the core entity terms facilitates the visualization and exploration of closely related evidence online.

6. Conclusion

We have presented a novel technique CCSE to retrieve those scholarly articles that share similar core contents. Two articles that share similar core contents should have similar research goals, backgrounds, and conclusions. Given an article a, retrieval of those articles that share similar core contents with a is a routine job that researchers strive to do to cross-validate the evidence already published in literature. As the terms related to the core content of an article may scatter in an article, it is challenging to recognize the core content and estimate inter-article similarity based on the core contents of articles. To tackle the challenge, CCSE is developed based on two hypotheses: (1) degrees of relatedness of a term t to the core content of an article a may depend on the positions where t appears in the title and the abstract of a; and (2) mismatch between any parts of the core contents of two articles a_1 and a_2 may significantly reduce the similarity between a_1 and a_2.

Performance of CCSE has been investigated using biomedical data in which core-content-sharing scholarly articles are selected by domain experts. CCSE performs significantly better than several state-of-the-art baselines, including text-based, link-based, and hybrid techniques. CCSE works on titles and abstracts of articles only, which are more publicly obtainable than bibliography that is employed by link-based and hybrid techniques. CCSE works for those articles that present research background followed by main results and discussion. It may be used to support the identification and validation of the closely related evidence already published in these articles.

Acknowledgement

This research was supported by the Ministry of Science and Technology (grant ID: MOST 105-2221-E-320-004) and Tzu Chi University (grant IDs: TCRPP103020 and TCRPP104010), Taiwan.

References

Aljaber, B., Stokes, N., Bailey, J., & Pei, J. (2010). Document clustering of scientific texts using citation contexts. *Information Retrieval*, *13*(2), 101-131.

Becker, K. G., Barnes, K. C., Bright, T. J., & Wang, S. A. (2004). The Genetic Association Database. *Nature Genetics,* 36(5), 431-432.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389-2404.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., et al. (2011). Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. *PLoS ONE,* 6(3), e18029.

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology*, *64*(9), 1759-1767.

Calado, P., Cristo, M., Moura, E., Ziviani, N., Ribeiro-Neto, B., & Goncalves, M. A. (2003). Combining Link-Based and Content-Based Methods for Web Document Classification. in *Proc. of the 2003 ACM CIKM International Conference on Information and Knowledge Management* (*CIKM'03*), New Orleans, Louisiana, USA.

Couto, T., Cristo, M., Gonçalves, M. A., Calado, P., Nivio Ziviani, N., Moura, E., et al. (2006). A Comparative Study of Citations and Links in Document Classification. in *Proc. of the 6th ACM/IEEE-CS joint conference on Digital libraries*, 75-84.

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, *59*(1), 51-62.

Gipp, B., & Beel, J. (2009). Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis. in *Proc. of the 12th International Conference on Scientometrics and Informetrics*, Brazil, 571-575.

Gipp, B., & Meuschke, N. (2011). Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence. in *Proc. of 11th ACM Symposium on Document Engineering*, Mountain View, CA, USA.

Glenisson, P., Glanzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing and Management, 41*, 1548-1572.

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

Janssens, F., Zhang, L., & De Moor, B. (2009). Glänzel W. Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management, 45*, 683-702.

Joachims, T. (2002). Optimizing Search Engines using Clickthrough Data. In *Proceedings of ACM SIGKDD*, Edmonton, Alberta, Canada, 133-142.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10-25.

Kumar, S., P. Reddy K., Reddy, V. B., & Singh, A. (2011). Similarity Analysis of Legal Judgments. in *Proc. of the Fourth Annual ACM Bangalore Conference* (*COMPUTE 2011*), Bangalore, Karnataka, India.

Landauer, T. K., Laham, D., & Derr, M. (2004). From paragraph to graph: Latent semantic analysis for information visualization. in *Proceedings of the National Academy of Sciences of the USA*, 101(Suppl 1), 5214-5219.

Lin, J., & Wilbur, W. J. (2007). PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, *8*, 423.

Liu, R.-L. (2015). Passage-based Bibliographic Coupling: An Inter-Article Similarity Measure for Biomedical Articles. *PLOS ONE, 10*(10), e0139245.

Liu, R.-L. (2017). Identification of Biomedical Articles with Highly Related Core Contents. *Proc. of the 9th Asian Conference on Intelligent Information and Database Systems*, 217-226, Kanazawa, Japan.

Liu, R.-L., & Huang, Y.-C. (2011). Ranker Enhancement for Proximity-based Ranking of Biomedical Texts. *Journal of the American Society for Information Science and Technology*, *62*(12), 2479-2495.

Liu, S., Chen, C., Ding, K., Wang, B., Xu, K., & Lin, Y. (2014). Literature retrieval based on citation context. *Scientometrics*, *101*(2), 1293-1307.

Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y., & De Moor, B. (2010). Weighted Hybrid Clustering by Combining Text Mining and Bibliometrics on a Large-Scale Journal Database. *Journal of the American Society for Information Science and Technology, 61*, 1105-1119.

Liu, X., Zhang, J., & Guo, C. (2013). Full-text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology*, *64*(9), 1852-1863.

Lu, Z., & Hirschman, L. (2012). Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database*, Vol. 2012, bas043.

Nakov, P. I., Schwartz, A. S., & Hearst, M. (2004). Citances: Citation sentences for semantic analysis of bioscience text. in *Proceedings of the SIGIR'04 workshop on search and discovery in bioinformatics*, 81-88.

PubMed. (2014). Computation of Related Citations. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Computation_of_Similar_Articl (accessed November, 2014).

Ritchie, A., Teufel, S., & Robertson, S. (2008). Using Terms from Citations for IR: Some First Results. in *Advances in Information Retrieval*, vol. 4956, Macdonald C, Ounis I, Plachouras V, Ruthven I, White R. (eds.), Springer, 211-221.

Robertson, S. E., Walker, S., & Beaulieu, M. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive. in *proceedings of the 7th Text REtrieval Conference* (TREC 7), Gaithersburg, USA, 253-264, 1998.

Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary

investigation. *Scientometrics*, *87*(2), 373-388.

Small, H. G. (1973). Co-citation in the scientific literature: A new measure of relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265-269.

Veloso, A., Almeida, H. M., Goncalves, M., & Meira Jr, W. (2008). Learning to Rank at Query-Time using Association Rules. In *Proceedings of the 31rd annual international ACM SIGIR conference on research and development in information retrieval, Singapore*, 267-274.

Whissell, J. S., & Clarke, C. L. A. (2013). Effective Measures for Inter-Document Similarity. in *proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (CIKM'13), 1361-1370.

Wiegers, T. C., Davis, A. P., Cohen, K. B., Hirschman, L., & Mattingly, C. J. (2009). Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD). *BMC Bioinformatics*, *10*, 326.

[About the authors]

Rey-Long Liu is currently a professor in Department of Medical Informatics, Tzu Chi University, Taiwan. He received his Ph.D. from Institute of Computer Science, National Tsing Hua University, Taiwan, 1994. He served as a professor and the chairman in Department of Information Management (Chung Hua University) and Department of Medical Informatics (Tzu Chi University). His research interest lies on intelligent information systems, with special focuses on text retrieval and mining as well as intelligent multi-agent systems, with application to biomedicine and business management. He has received many awards in teaching, research, and guidance counseling. He also served as program committee and/or editorial board members for many internationally renowned publications, including *ACIIDS*, *PRIMA*, and *IP&M*.