Future Generation Computer Systems 26 (2010) 400–407

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

A multiple-perspective approach to constructing and aggregating Citation Semantic Link Network☆

Zhixing Huang*, Yuhui Qiu

College of Computer and Information Science, Southwest University, Tiansheng Road #2, Beibei Distinct, Chonqing, 400715, PR China

ARTICLE INFO

Article history:
Received 22 December 2008
Received in revised form
3 June 2009
Accepted 24 July 2009
Available online 3 August 2009

Keywords:
Semantic Link Network
Opinion mining
Sentiment analysis
Community identification

ABSTRACT

Various kinds of semantic relationships exist among scientific literatures which worth to be explored. This paper proposes a Citation Semantic Link Network (C-SLN) to describe the semantic information over the literature citation networks. A framework of the construction of C-SLN is represented by integrating several NLP methods. The methods of aggregating a C-SLN and the algorithms of discovering opinion communities in a C-SLN are also discussed. Based on a multi-perspective exploration on the C-SLN, we can effectively find articles of high importance, aggregate the function of citations and detect opinion communities among scientific documents.

1. Introduction

Understanding the existing research papers lays a foundation for scientists to develop their own research agendas, communicate with others, and keep up with up-to-date research. When a scientific article is written, the author often cites previous articles in support of his own arguments; those articles, whereas, have cited other papers for the same purpose. As time passes, these cited relations between articles form a network. The citation network enables researchers to find the chronological order and the progress in a particular field of study.

Historically, bibliometric metrics have been commonly used to measure the impact of a researcher's work by how often they are cited. However, just knowing that an article cited rate is often not enough. Researchers from the field of discourse studies have pointed out that many citations are perfunctory [1] or done out of "politeness, policy or piety" [2]. Keyword-based methods have been successfully used in retrieving documents automatically. Scientific literature search engines like CiteSeer[1] and Google Scholar[2] can help search research papers and provide scientists with important information such as citations and citation context for users to browse. Browsing each citation is useful, but time-consuming. Experienced researchers are often interested in detailed relations among articles. They want to know if a certain article criticizes another and what the criticism is, or if the current work is based on that prior work [3]. This type of information is hard to come by using current search technologies. Neither the author's abstract, nor raw citation counts help users in assessing the relation between articles [4]. Not all search needs are fulfilled by current citation indexers.

Studies have shown that different citations to the same article often focus on different aspects of that article. There are many important information such as comparison, relatedness, and argumentation that can be drawn through the citation relationship. To explore those semantic information, we need a formal representation framework based on which we can do more searches and queries. A recent trend about analyzing the relationship between documents is focusing on content mining and its strong connection with natural language processing (NLP) [5]. Applying NLP techniques into scientific document content analysis is becoming increasingly necessary.

In this paper, we present a formal knowledge representation framework Citation Semantic Link Network (C-SLN) to describe the citation semantics among scientific literatures. Based on the C-SLN, one can apply various statistic methods to further explore the semantic information from different perspectives. C-SLN enables users to build up a more expressive citation network among literatures. It can be used not only to support basic reasoning and queries, but also to discover deep semantic relationships among literatures.

☆ This work was partially supported by National Basic Research Program of China (973 project no. 2003CB317008), Natural Science Foundation Project of CQ CSTC (no. 2008BB2005) and Scientific Research Foundation of Southwest University (no. SWUB2007056).

* Corresponding author.
E-mail addresses: huangzx@swu.edu.cn (Z. Huang), yhqiu@swu.edu.cn (Y. Qiu).

1 http://citeseer.ist.psu.edu/.

2 http://scholar.google.com/.

0167-739X/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2009.07.006

The contribution of this paper includes the following four aspects: (1) We propose a general framework of constructing C-SLN which extends the semantics of typical citation network by combining several NLP techniques such as citation function classification, sentiment analysis and keyword extraction, etc.; (2) We propose some measures to aggregate multiple citations of a reference from intra-article and inter-article perspectives. (3) The algorithms of discovering disjoined and overlapping opinion communities in C-SLN are discussed. (4) A concrete case of C-SLN in opinion mining discipline is studied.

2. Related work

Utilizing rich semantic relationships among versatile resources can assist us to accomplish complex tasks and solve problems in the future interconnection environment [6–8]. The Semantic Link Network (SLN) is a semantic data mode for organizing various resources by attaching semantic factors on links. The chapter 2 of *The Knowledge Grid* systematically introduces the fundamental concepts, operations and properties of SLN as well as a series of normal forms, criteria, constraints to guarantee SLN completeness and consistency [9]. SLN has been used to improve the efficiency of query routing in P2P network [10], object prefetching [11] and it has been adopted as one of the major resource organization mechanisms for the Knowledge Grid [9,12]. Although there exist some software tools such as SLN-Builder [13] that enable definition, modification and verification of, as well as access to the SLN manually, one of the main challenges in realizing the SLN lies in effective approaches to automatical discovering semantic links from various resources according to some definitions on concepts and rules in the primitive semantic space [14].

Citation Semantic Link Network can be regarded as an instance of SLNs by attaching semantic factor on citation links. Various approaches in citation analysis have been applied to extend the capability of the traditional citation network. Zhang and Koppaka described the use of a semantics-based citation network in a legal research tool. To distinguish citations based on legal issues, they use the concept of Reason for Citing (RFC), and dissect the general multi-dimensional network into subnetworks. The system allows legal professionals to efficiently study legal issues without having to go through the whole cases or tedious manual citation search [15]. Aleman-Meza et al. proposed a semantic web application that detects Conflict of Interest (COI) relationships among potential reviewers and authors of scientific papers [16]. Qazvinian et al. proposed a model of summarizing a single article, which is based on analyzing others' viewpoint of the target article's contributions and the study of its citation summary network using a clustering approach [17]. Our work extends the existing methods by integrating several natural language processing techniques, including citation function classification [4,18], sentiment analysis [19–21] and keyword extraction, to construct citation Semantic Link Networks.

Detecting and identifying community is an essential issue for social network as well as semantic network. Palla et al. introduced an approach to analyzing the statistical features of the interwoven sets of overlapping *k*-clique communities [22]. Farkas et al. proposed a clique percolation method with weights for weighted networks based on the concepts of percolating *k*-cliques with high enough intensity [23]. Bansal and Chawla analyzed the clustering problem that given a complete graph with edges labeled positive or negative, finds a partition of the vertices into clusters that agrees as much as possible with the edge labels [24]. Yang proposed an algorithm to mine signed social networks with positive and negative links [25]. Zhuge proposes three types of approaches to discovering reasoning-constraint, rule-constraint and classification-constraint semantic communities in SLN [14], and a novel community discovery method by using the topological centrality [26].

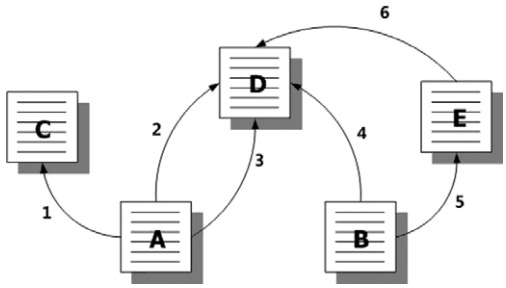

Fig. 1. An example of citation network.

3. General architecture

Before proposing our framework of constructing Semantic Link Networks among scientific papers, let us briefly outline some important characteristics which are often ignored in typical citation networks.

1. Frequency: A reference can be cited in an article more than one time (An example of citation network is shown in Fig. 1). Among those references in a given paper, some references may be cited more frequently than others, since their contents are more relevant to the current article.
2. Multi-dimension: The citations of the same reference may focus on the same or different aspects of the reference. An important fact is that a reference can be cited many times and each citation may play a different role in a scientific article.
3. Location: The location of a sentence containing a reference can partially reflect the relevance between the articles. The citation appearing at the beginning of the paper may have a different status compared with those towards the end of it.
4. Opinion and Sentiment: Authors often express their opinions about references when they cite them. The expression with positive, neutral or negative orientation can be explicitly or implicitly found.

From the above discussion, we can find that those aspects are important when analyzing the relationship among scientific documents and valuable for constructing a semantic-based citation network.

3.1. Tasks

Fig. 2 depicts the general architecture for creating a semantic-based citation network. There are five technical tasks need to be performed:

1. Divide scientific articles into different parts. A scientific document generally includes abstract, introduction, body, experiment, conclusion and reference parts.
2. Extract consecutive sentences, called citation context, in each part according to the article architecture and citations through context analysis. Each citation context may contain one or more citation instances.
3. Classify the function of each citation instance and identify the sentiment orientation that the author has expressed his (positive, neutral or negative) opinions on. For each citation instance, extract the keywords which can be used to describe the main topic of the citation.
4. Organize the citation instances into a C-SLN, calculate the aggregate information among links after checking the correctness of the citation links through modification and visualization.
5. Deploy the C-SLN and explore emergent semantics using semantic reasoning rules. The C-SLN can be future utilized in support of other data models, such as Knowledge Flow Model [9] and Resource Space Model [27].

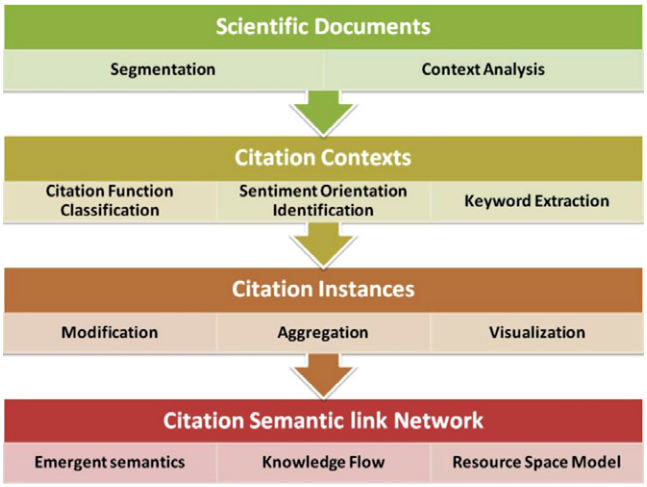

Fig. 2. The general architecture for constructing a C-SLN.

In the following, we will mainly focus on formal presentation, construction and aggregation aspects of C-SLN. For simplicity, NLP techniques adopted in this article, such as segmentation and context analysis, will not be exhaustively discussed.

4. Citation Semantic Link Network

A C-SLN is a directed acyclic graph (DAG) consisting of semantic nodes and semantic links between nodes. A semantic node in C-SLN represents a scientific article and has the attributes such as title, author, publication date and length. A semantic link directed from one node (predecessor) to another (successor) can be represented as a pointer labeled with a semantic property. The C-SLN is a special kind of SLN, for its semantic links generally represent the reference relationships between the semantic nodes.

Definition 1. Let $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ be a set of scientific articles. Each article α_i has a set of references $\mathcal{R}_i = \{\gamma_{i,1}, \gamma_{i,2}, \ldots, \gamma_{i,n_i}\}$, where n_i represents the number of references in article α_i.

Definition 2. Suppose a reference $\gamma_{i,j} \in \mathcal{R}_i$ has been cited $m_{i,j}$ times in article α_i, where $1 \leq j \leq n_i$ and $m_{i,j} > 0$. We let the citation set of reference $\gamma_{i,j}$ be $\mathcal{C}_{i,j} = \{c_{i,j}^1, c_{i,j}^2, \ldots, c_{i,j}^k, \cdots, c_{i,j}^{m_{i,j}}\}$, where $c_{i,j}^k$, $1 \leq k \leq m_{i,j}$, represents the kth citation of reference j in article α_i.

Definition 3. A citation semantic link represents a directed relationship between two nodes, denoted by $X \xrightarrow{c} Y$, where X and Y are the nodes, and $\xrightarrow{c}$ is called a c-link.

A c-link can be represented as a 7-tuple $\langle so, ta, ca, op, ke, lo, an\rangle$, where so represents the source article, ta represents the reference article cited by so, ca represents the category of citation function, op represents author's opinion orientation about the citation, ke represents the keywords in the citation context, lo represents the location of a citation, and an represents the context sentences of a citation.

4.1. Construction of C-SLN

Three natural language processing techniques are adapted for constructing a C-SLN in this paper. They include citation function classification, sentiment analysis and keyword extraction.

Citation function is defined as the author's reason for citing a given paper [4,28]. Different from [4] what uses a flat, one-dimensional classification scheme, we separate sentiment classification task from the citation function classification process. Moreover, some categories discussed in [4] which are hard to identify automatically (or even manually) are merged into corresponding general categories in this paper. The categories are as follows:

- Compare/Contrast: Contrast or comparison with other work.
- Similar: Author's work and cited work are similar.
- Use: The contribution in the cited work is used by the current work, e.g. *"We used WordNet [x] to find their synonyms."*
- Weak: Explicitly state the weakness of the reference. For instance, *"Exiting techniques in [x] are not suitable for this case."* The article addresses the weakness of the reference *x*.
- Detail: Reference provides more detailed information. It is used in the case of space limitation or issues out of the scope of the paper. A typical sentence of this category likes this: "*See [x] for details.*"
- Mention: Neutral description of cited work, or not enough textual evidence for above categories.

Since there is a set of verbs and sentence patterns that has been frequently used in each citation category, these verbs and patterns can be used as the cues for automatic citation function classification. Similar to [4], we adapt a supervised learning method for automatic citation function classification. In the training phase, the typical cue verbs and sentence patterns are manually extracted and classified into the corresponding categories. In the learning phase, a citation belongs to a category unless the explicit cue verb or sentence pattern is matched.

Now, we will discuss the task of identifying authors' attitudes (positive, neutral or negative) towards the works they cited in their articles. See the following examples taken from [20]:

(a) Mooney and Bunescu (2005) gave a *good* survey and comparison of existing methods. Conditional random field (CRF) (Lafferty, McCallum and Pereira 2001) is perhaps the *best* method so far.
(b) Researchers in linguistics have studied the syntax and semantics of comparative constructs for a long time (e.g., Moltmann 1997; Doran et al. 1994; Kennedy 2005). *However*, they have *not* examined computational methods for extracting comparative sentences and relations.
(c) In (Hu and Liu 2004), some methods were proposed to extract opinions from customer reviews.

Using sentiment analysis methods, citations (a), (b) and (c) can be identified to be positive, negative and neutral respectively, since evidential sentiment words *"good"* and *"best"* (positive) are found in (a), *"however" and "not"* (negative) are found in (b) and none is found in (c). Various semantically annotated lexical tools (such as SentiWordNet [29], OpinionFinder [30], General Inquirer [31], etc.) can help us automatically identifying author's sentiment orientation in each citation.

Moreover, the keywords associated with citation links contain the fundamental semantic information of the citation. They represent which aspect the citation focused on. The keywords can be extracted in citation context by adopting some information extraction techniques such as TFIDF algorithm [32].

Integrating these semantics (i.e. citation function classification, sentiment analysis and keyword) into citation networks can provide more informative evidences, which can help us discover documents without having to go through whole corpus or tedious manual citation search.

4.2. Represent C-SLN in XML

There are two main types of classes in C-SLN, one class representing the articles and the other representing the citation links. The article information includes title, id, author name(s), abstract, keywords, publication source, month, year and address, etc. The information of an article can be depicted in XML format as in the following style:

```
<?xml version="1.0" encoding="UTF-8"?>
<article>
    <id> Paper's Id</id>
    <title> Paper's Title </title>
    ...
</article>
```

Moreover, we can represent a citation semantic link in XML format. For instance, an article A cites article B, C, and D in its introduction section with the following sentence (digested from [21]):

"*The most existing methods for detecting web spam and email spam [B, C, D] are unsuitable for review spam.*"

A citation link $A \xrightarrow{c} B$ can be represented in XML as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<citation>
    <source> A.id </source>
    <target> B.id </target>
    <category> Criticize </category>
    <sentiment> Negative </sentiment>
    <feature> web spam, email spam,
    review spam </feature>
    <location> Introduction </location>
    <annotation>
    The most existing methods for
    detecting web spam and email spam
    [B,C,D] are unsuitable for review
    spam.
    </annotation>
</citation>
```

Analogously, the citation links $A \xrightarrow{c} C$ and $A \xrightarrow{c} D$ can be represented and stored in XML file as the same way.

4.3. Aggregation

Each citation can be described as a *c-link* as defined in Definition 2. Moreover, for the sake of simplicity, we let $t^k_{i,j}$, $o^k_{i,j}$, $F^k_{i,j}$, and $l^k_{i,j}$ represent the category, opinion orientation, feature properties of citation $c^k_{i,j}$ respectively.

4.3.1. Intra-article analysis

We can calculate the importance of a reference $\gamma_{i,j} \in R_i$ within article α_i:

$$I(r_{i,j}) = \frac{\sum_{c^k_{i,j} \in C_{i,j}} W_t(t^k_{i,j}) \times W_l(l^k_{i,j})}{\max_{1 \le i \le n_i} \left\{ \sum_{c^k_{i,j} \in C_{i,j}} W_t(t^k_{i,j}) \times W_l(l^k_{i,j}) \right\}}. \quad (1)$$

W_t and W_l represent the weight function of category and location individually. Intuitively, a reference appears many times in the main part of an article should have a higher importance than that of the references appear once in other parts.

In the simplest case, if we let all the weights be 1. That is:

$$I(\gamma_{i,j}) = \frac{m_{i,j}}{\max_{1 \le j \le n_i} \{m_{i,j}\}}. \quad (2)$$

Aggregating author i's overall opinion orientation of reference $\gamma_{i,j}$ can be calculated by averaging of all the $o_{i,j}$ in the article. Hence, we have:

$$O(\gamma_{i,j}) = \frac{\sum_{1 \le k \le m_{i,j}} o^k_{i,j}}{m_{i,j}}. \quad (3)$$

Without loss of generality, we assume that all the $o^k_{i,j}$ are drawn from set $\{1, 0, -1\}$. The elements represent *positive, neutral,* and *negative* respectively. Thus, the aggregated opinion orientation $O(\gamma_{i,j})$ is in the interval $[-1, 1]$.

The function of citations are non-numerical, however they can be aggregated based on their strengths of the reason for citing. For example, a citation function aggregation rule can be described as in the following:

$$Compare > Detail > Similar > Use > Weak > Mention.$$

The citation function of the highest strength is kept in the aggregation process.

Aggregating multiple citations of a reference into a single citation link makes it easy to understand the role of the reference and to visualize a large-scale C-SLN. In addition, two citations cite the same reference may focus on different aspects of the article. To determine whether two citation links focus on the same aspect, a feasible and effective method is to calculate the semantic distance between the keywords or annotations attributes. The keywords also can be used to reduce the scale of a C-SLN by filtering the links with low semantic similarity.

4.3.2. Inter-article analysis

Definition 4. Suppose $\mathcal{R} = \mathcal{R}_1 \cup \mathcal{R}_2 \cup \cdots \cup \mathcal{R}_N$ represents the overall reference set of article set $\mathcal{A}$. And $\mathcal{SR}(\tau)$, $\mathcal{SR}(\tau) \subseteq \mathcal{A}$, represents the article set which has cited reference τ and $\lambda(\tau)$ represents the number of articles in set $\mathcal{SR}(\tau)$.

For any reference $\tau \in \mathcal{R}$, The importance of τ is the aggregation of its importance among all different articles. We define that $I(\tau, i) = I(\gamma_{i,j})$ if there exists a reference $\gamma_{i,j} = \tau$, else $I(\tau_i) = 0$, thus we have:

$$\bar{I}(\tau) = \sum_{1 \le i \le N} I(\tau, i). \quad (4)$$

Suppose $O(\tau, i)$ represents the opinion orientation of article α_i about reference τ. $O(\tau, i) = O(\gamma_{i,j})$ if α_i in $\mathcal{SR}(\tau)$, else $O(\tau, i) = 0$. Then, the overall authors' opinion orientations of reference τ can be calculated as:

$$\bar{O}(\tau) = \frac{\sum_{\alpha_i \in \mathcal{SR}(\tau)} O(\tau, i)}{\lambda(\tau)}. \quad (5)$$

Aggregating those information, we can easily identify the articles of high importance and detect the articles of high criticizing or approving rate.

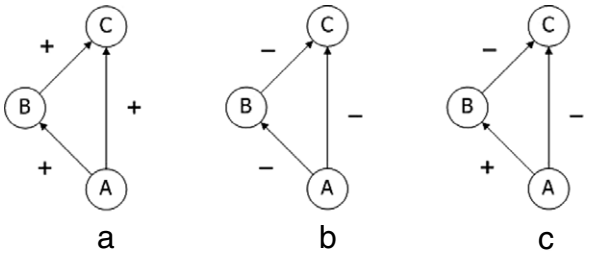

Fig. 3. Opinion community.

4.4. Discover opinion communities

Since different articles may have different opinions on the same reference, discovering opinion communities in C-SLNs can help us group those documents. We assume that each paper has one opinion (positive or negative) about a reference in a certain aspect after the aggregation calculation process. Thus, the relationship between articles can form a citation network. Each pair of the network has one directed edge between them. The edges are labeled with "+" or "−" to represent authors' positive or negative opinion about the linked reference.

Definition 5 (*Disjointed Opinion Community*). For the node set $\mathcal{C}$ of a C-SLN, it can be split into disjointed opinion communities $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_s$, and the communities satisfy that:

(1) $\mathcal{C} = \bigcup_{1 \le k \le s} \mathcal{C}_k$;
(2) For any community pair $(\mathcal{C}_p, \mathcal{C}_q)$, where $1 \le p, q \le s$, has $\mathcal{C}_p \cap \mathcal{C}_q = \emptyset$;
(3) For any node pair (i, j) within $\mathcal{C}_k$, if there exists a link c between node i and j, then $c.op$ must be non-negative;
(4) For any community pair $(\mathcal{C}_p, \mathcal{C}_q)$, either there exist a node pair (i, j), where $i \in \mathcal{C}_p$ and $j \in \mathcal{C}_q$, $c.op$ is negative, or for all the node pairs there is no link existing.

Algorithm 1 *DisjointedCommunity*(*LinkedList* : *seq*)

```
1:  s = 0; i = seq.getFirst();
2:  while seq.hasNext() = true do
3:    for k = 1 to s do
4:      compatible = false;
5:      if i is compatible with C_k then
6:        C_k = C_k ∪ {i};
7:        compatible = true;
8:        break;
9:      end if
10:   end for
11:   if compatible = false or s = 0 then
12:     s = s + 1;
13:     C_s = {i};
14:   end if
15:   i = seq.next();
16: end while
17: return
```

We take some examples to demonstrate this definition. In Fig. 3, (a) forms one community $\{A, B, C\}$; (b) can be split into three communities $\{A\}\{B\}\{C\}$; and (c) can be split into two communities $\{A, B\}\{C\}$. In general, the partition of disjoined opinion community of a C-SLN is not always unique. See the example shown in Fig. 4. Network (a) can be split into sets $\{A, B, C\}\{D\}$ or $\{A, C, D\}\{B\}$ as demonstrated in (b) and (c) respectively. One of the solutions can be found by using Algorithm 1.

In Algorithm 1, parameter *seq* represents one of the topological sort of the C-SLN. An topological sort is a linear ordering of its nodes in which each node comes before all nodes to which it has outbound edges. Node i is compatible with $\mathcal{C}_k$ means that the links between i and the nodes in $\mathcal{C}_k$ are all non-negative. As shown in line 5, node i is appended into one compatible C_k. It is easy to conclude that each topological sort maps one partition result.

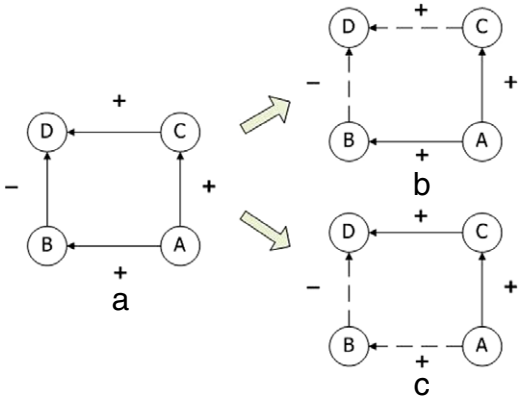

Fig. 4. Different partitions.

If different communities are allowed to sharing the common node or nodes, we can find overlapping communities of a C-SLN. The overlapping opinion community is defined as follows:

Definition 6 (*Overlapping Opinion Community*). For the node set $\mathcal{C}$ of a C-SLN, it can be split into overlapping opinion communities $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_s$, and the communities satisfy that:

(1) $\mathcal{C} = \cup_{1 \le k \le s} \mathcal{C}_k$;
(2) For any communities $\mathcal{C}_p$ and $\mathcal{C}_q$, $\mathcal{C}_p \ne \mathcal{C}_q$;
(3) For any node pair (i, j) within $\mathcal{C}_k$, if there exists a link c between node i and j, then $c.op$ must be non-negative;
(4) For any community pair $(\mathcal{C}_p, \mathcal{C}_q)$, either there exists a node pair (i, j), where $i \in \mathcal{C}_p$ and $j \in \mathcal{C}_q$, $c.op$ is negative, or for all the node pairs there is no link existing.

The overlapping opinion community consists of the articles which share similar or non-conflicting opinion. An example of overlapping communities is shown in Fig. 5. The nodes can be grouping into two overlapping communities $\{A, B\}$ and $\{B, C\}$ with the common node $\{B\}$. Algorithm 2 is designed to discover such overlapping communities, with a slight modification of Algorithm 1. Different from the previous algorithm, node i is appended into all the compatible communities.

Since some disjoined or overlapping opinion communities may be very large in a C-SLN, these communities can be further split into smaller communities by using other community discovery methods. For instance, we can find some small k-clique[3] communities in each opinion community by using CPM [22] algorithm. These new communities represent the articles which are highly related with each other and share similar or non-conflicting opinion.

Discovering opinion communities can help us find the authors who have conflicting research opinions and who are actively in argument with other authors.

[3] A k-clique community is a union of all k-cliques (complete subgraphs of size k) that can be reached from each other through a series of adjacent k-cliques (where adjacency means sharing $k - 1$ nodes).

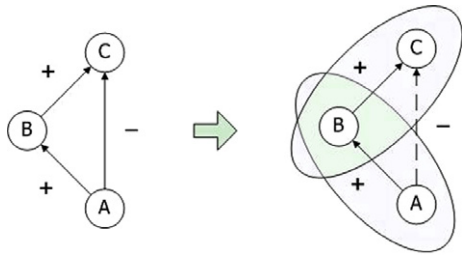

Fig. 5. Overlapping community.

Algorithm 2 *OverlappingCommunity*(*LinkedList* : *seq*))

```
1: s = 0; i = seq.getFirst();
2: while seq.hasNext() = true do
3:   for k = 1 to s do
4:     compatible = false;
5:     if i is compatible with C_k then
6:       C_k = C_k ∪ {i};
7:       compatible = true;
8:     end if
9:   end for
10:  if compatible = false or s = 0 then
11:    s = s + 1;
12:    C_s = {i};
13:  end if
14:  i = seq.next();
15: end while
16: return return
```

4.5. Visualization and modification

After constructing and aggregating C-SLN automatically, we can visualize or modify it with the software tool SLN-Builder [13]. We also can find abnormal nodes or edges in a semantic graph to ensure the correctness and effectiveness of C-SLN. The annotation associated with each semantic link provides the necessary context information about the citation. The information is useful when we check the correctness of the link in C-SLN.

4.6. Discover emergent semantic topics

The reasoning rules in SLNs also can be employed in C-SLNs. For example, if article *a* is similar to *b*, and *b* is similar to *c*, then *a* is similar to *c*. More reasoning rules about SLNs are discussed in [9,14]. In addition, based on the citation Semantic Link Network, we can find more emergent semantic topics. Here are some heuristic rules:

1. For a novice in a research domain, he or she should read the article with high inter-article importance at first before doing further study and research.
2. If the two papers compare the same article, the methods proposed in these two papers also need to be compared. It can be one of the research topics or future work for the authors with the same research interests.
3. If an article has been frequently cited and identified as the *Use* citation category, the method proposed in this article must be very popular and powerful.
4. With time passing by, new nodes and new links will be appended into C-SLN. Similar to the method proposed in [33], by comparing C-SLNs in different time intervals, e.g. in years, we can find emergent and identifying emergent semantic nodes. Meanwhile, by analyzing the keywords associated with semantic links and the information stored in the nodes, we can discover the newly emergent research topics. Sometimes, a new semantic community may become a new research discipline.

Table 1
Percentage of each category.

Category	Number	Percentage
Compare	31	0.10
Weak	24	0.08
Detail	6	0.02
Mention	203	0.67
Similar	3	0.01
Use	38	0.12

Table 2
Percentage of each location.

Location	Number	Percentage
Introduction	75	0.24
Related work	137	0.35
Body	90	0.29
Experiment	1	<0.01
Conclusion	2	<0.01

These heuristic rules can help us effectively understand the semantics behind complex citation networks.

5. Experiment

To explicitly demonstrate the function of our system, we select the corpus with a relatively small amount of articles.

5.1. Data set

Our evaluation corpus for citation analysis consists of 10 articles from Prof. Bing Liu[4] and his colleagues. The articles are their recent published papers including two fields: Opinion Mining and Visual Data Mining. Each citation instance is automatically or semi-automatically extracted and tagged with its corresponding citation type. Then, a C-SLN is constructed based on those citation instances.

5.2. Experiment results

The 10 articles contain a total of 305 citation instances. The percentage of each type and the percentage of each location are described in Tables 1 and 2 respectively.

As shown in Table 1, we can find that the *Similar* category is less than 1%. This happens in cases where authors do not want to admit (or stress) that their works are similar to somebody else's method. In addition, about two-thirds of citations belong to *Mentioned* category. Table 2 confirms the commonsense that as a good writing skill, an author should introduce related articles of his work in the *Related Work* section. Table 2 also demonstrates that the authors like to cite related works in the *Introduction* and the main body sections of their articles. Moreover, the citations are relatively rare in *Experiment* and *Conclusion* sections.

Table 3 shows the percentages of authors' sentiment orientation of different citation categories. In most cases, the authors are not likely to explicitly express their positive sentimental orientations. As shown in Table 3, less than 10% cases are cited with explicit positive orientation expression.

Since a C-SLN may contain many nodes and edges (arcs), we should visualize and navigate the C-SLN by filtering out those less important elements. Fig. 6 demonstrates a view of the C-SLN. The view is constructed by selecting three most important references of each article in evaluation corpus. The graph includes 27 nodes and

[4] Professor Bing Liu is from the University of Illinois at Chicago (UIC). His homepage is at http://www.cs.uic.edu/liub/.

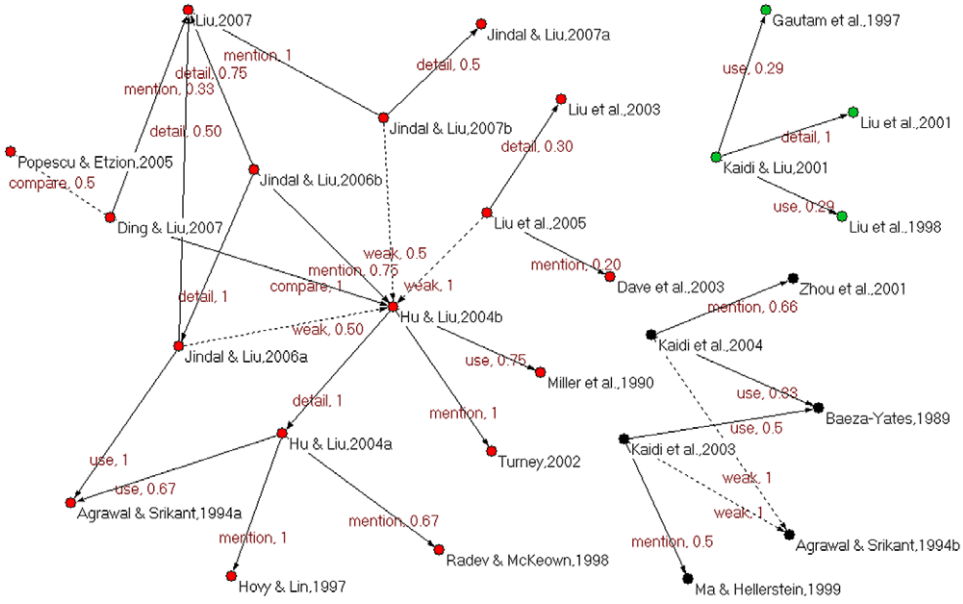

Fig. 6. A View of a C-SLN.

Table 3
Sentimental orientation in each category.

Category	Negative	Neutral	Positive
Compare	16	12	3
Weak	24	0	0
Detail	0	5	1
Mention	9	179	16
Similar	0	3	0
Use	0	38	0
Total	49 (16%)	236 (77%)	20 (7%)

30 arcs. The label associated with each arc indicates the aggregated intra-article importance and sentiment orientation as we have discussed in Section 4. The graph is composed of three disjoined subgraphs (semantic communities). The largest part describes the citation relationship of the articles in opinion mining field, and the small two parts describe the citations in visual data mining field. The dash lines indicate *negative* links, while the solid line indicates positive or neutral links, the label associated with links represent corresponding citation function and intra-importance.

Different from traditional citation network, an article with the highest citation frequency may not be the most important article in the corpus. We can find that the article "Hu and Liu, 2004a" is in the center of these documents. Intuitively, it might be the most important article in this corpus. When we list the most important references through inter-article analysis and community identification, the result confirms that "Hu and Liu, 2004a" is the most important article in the corpus. In contrast, The article "Dave et al., 2003" with the highest citation frequency is just ranked as the 4th important article. The view of a C-SLN can help users to explore the comprehensive relationship among documents. For example, for a novice who wants to research in opinion mining discipline, a fitting recommendation is that he (she) should read the scientific articles with high importance at first. Moreover, the significant contributions are often associated with "*Use*" links. For instance, article "Agrawal and Srikant, 1994a" which proposed an approach for mining sequential patterns are with two "*Use*" links.

Fig. 6 also demonstrates a view of authors' sentimental orientations in their articles. It is quite interesting that the article with the high inter-article importance is often associated with a large amount of negative citations. For example, article "Hu and Liu, 2004b" has 3 negative in-edges and 2 neutral in-edges. Indeed, this article has been cited totally 21 times in the corpus and 7 citations among them are with negative orientation. It is true that a good scientific article can inspire and motivate other researchers searching for new methods to solve a certain problem. By retrieving the annotation associated with negative links, we can find the limitations of a given reference.

6. Conclusion and future work

In this paper, we propose an approach to constructing citation semantic link network. It extends semantics of typical citation network by integrating several natural language processing techniques such as citation function classification, sentiment analysis and keyword extraction. By exploring C-SLN from the multiple perspectives, we can effectively understand the complex citation relationships among scientific documents. The proposed method is helpful in discovering the articles of high importance, aggregating the function of the citations and detecting the opinion communities.

The following issues need further efforts: efficient approaches to automatically constructing a C-SLN of a large-scale; effective mechanisms to analyze and visualize the evolution of C-SLN communities for detecting and identifying emergent topics; exploring the connections among C-SLN and other data models such as Knowledge Flow Model and Resource Space Model.

Acknowledgments

The authors would like to thank the editors of the special issue and the anonymous reviewers for providing many invaluable comments that improved this paper. Special thanks are due to Dr. Xiaoping Sun and Dr. Junsheng Zhang for their useful discussions. We thank Song Wu, Gang Liu, and Peng Xiao for their help of system prototype implementation.

References

[1] M.J. Moravcsik, P. Murugesan, Some results on the function and quality of citations, Social Studies of Science 5 (1975) 88–91.
[2] J.M. Ziman, Information, communication, knowledge, Nature 224 (1969) 318–324.
[3] S.B. Shum, Evolving the web for scientific knowledge: First steps towards an hci knowledge web, Interface, British HCI Group Magazine 39 (1998).

[4] S. Teufel, A. Siddharthan, D. Tidhar, Automatic classification of citation function, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, Sydney, Australia, 2006, pp. 103–110.
[5] B. Liu, From web content mining to natural language processing (tutorial), in: 45th Annual Meeting of the Association for Computational Linguistics, ACL-2007, Prague, Czech Republic, June 23–30, 2007.
[6] H. Zhuge, Semantics, resource and grid, Future Generation Computer System 20 (1) (2004) 1–5.
[7] D.D. Roure, C. Goble, R. Stevens, The design and realisation of the virtual research environment for social sharing of workflows, Future Generation Computer Systems 25 (5) (2009) 561–567.
[8] T. McPhillips, S. Bowers, D. Zinn, B. Ludäcker, Scientific workflow design for mere mortals, Future Generation Computer Systems 25 (5) (2009) 541–551.
[9] H. Zhuge, The Knowledge Grid, World Scientific Publishing Co., Singapore, 2004.
[10] H. Zhuge, X. Li, Peer-to-peer in metric space and semantic space, IEEE Transactions on Knowledge and Data Engineering 19 (6) (2007) 759–771.
[11] A.P. Pons, Object prefetching using semantic links, SIGMIS Database 37 (1) (2006) 97–109.
[12] H. Zhuge, Autonomous Semantic Link Networking model for the knowledge grid, Concurrency and Computation: Practice and Experience 7(19) (2007) 1065–1085.
[13] H. Zhuge, R. Jia, Semantic link network builder and intelligent browser, Concurrency and Computation: Practice and Experience 16 (14) (2004) 1453–1476.
[14] H. Zhuge, Communities and emerging semantics in Semantic Link Network: Discovery and learning, IEEE Transactions on Knowledge and Data Engineering 21 (6) (2009) 785–799.
[15] P. Zhang, L. Koppaka, Semantics-based legal citation network, in: Proceedings of the 11th international conference on Artificial intelligence and law, ICAIL'07, New York, NY, USA, 2007, pp. 123–130.
[16] B. Aleman-Meza, M. Nagarajan, L. Ding, A. Sheth, I.B. Arpinar, A. Joshi, T. Finin, Scalable semantic analytics on social networks for addressing the problem of conflict of interest detection, ACM Transactions on Web 2 (1) (2008) 1–29.
[17] V. Qazvinian, D. Radev, Scientific paper summarization using citation summary networks, in: International Conference on Computational Linguistics, COLING, Manchester, UK, 2008.
[18] I. Spiegel-Rosing, Science studies: Bibliometric and content analysis, Social Studies of Science 7 (1) (1977) 97–113.
[19] M. Hu, B. Liu, Mining opinion features in customer reviews, in: Proceedings of Nineteenth National Conference on Artificial Intellgience, AAAI-2004, San Jose, USA, 2004.
[20] N. Jindal, B. Liu, Mining comparative sentences and relations, in: The Eighteenth Innovative Applications of Artificial Intelligence Conference, Boston, USA, 2006.
[21] B. Liu, M. Hu, J. Cheng, Opinion observer: analyzing and comparing opinions on the web, in: Proceedings of the 14th international conference on World Wide Web, WWW05, ACM Press, New York, NY, USA, 2005, pp. 342–351.
[22] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814–818.
[23] I. Farkas, D. Ábel, G. Palla, T. Vicsek, Weighted network modules, New Journal of Physics 9 (3) (2007) 180–198.
[24] N. Bansal, A. Blum, S. Chawla, Correlation clustering, Machine Learning 56 (1–3) (2004) 89–113.
[25] B. Yang, W. Cheung, J. Liu, Community mining from signed social networks, IEEE Transactions on Knowledge and Data Engineering 19 (10) (2007) 1333–1348.
[26] H. Zhuge, J. Zhang, Topological centrality and its applications, CoRR abs/0902.1911, 2009.
[27] H. Zhuge, The Web Resource Space Model, Springer, 2008.
[28] A. Siddharthan, 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 2007, Rochester, New York, 2007, pp. 316–323.
[29] A. Esuli, F. Sebastiani, Sentiwordnet: A publicly available lexical resource for opinion mining, in: Proceedings of the 5th Conference on Language Resources and Evaluation, LREC06, 2006, pp. 417–422.
[30] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: HLT'05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 347–354.
[31] M.S. S. D. M.O. Philip J.Stone Dexter C.Dunphy, The General Inquirer: A Computer Approach to Content Analysis, The MIT Press, 1966.
[32] T. Joachims, A probabilistic analysis of the rocchio algorithm with tfidf for text categorization, in: ICML'97: Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, CA, USA, 1997, pp. 143–151.
[33] G. Palla, A. lászló Barabási, T. Vicsek, Quantifying social group evolution, Nature 446 (2007) 664–667.

Zhixing Huang received his Ph.D. in Artificial Intelligence from Southwest University, China, in 2006. He is an associate professor in the College of Computer and Information Science at Southwest University, China and is currently working as a postdoctoral researcher under the lead of Prof. Hai Zhuge at Knowledge Grid Research Group, Beijing, the Key Lab of Intelligent information Processing of Chinese Academy of Sciences. From Oct. 2006 to Sept. 2007, he was a visiting researcher in Ishida & Matsubara Lab of the Department of Social Informatics at Kyoto University, Japan. His research interests include semantic grid, cooperative computing and Information economics.

Yuhui Qiu, Professor of the Faculty of computer and information science in Southwest University (SWU), China, Ph.D. student adviser. Director of Institute of Artificial Intelligence, SWU; Member Of Professional Association; Vice-Director of Chinese Association of Artificial Intelligence; IEEE Senior Member; ACM Professional Member. His research interests include discrete mathematics, compiler principle, logical foundation of artificial intelligence, artificial intelligence, expert systems, automatic reasoning, machine learning, intelligent agents and computing intelligent Etc AI, DAI, and mobile computing.