Citation Role Labeling via Local, Pairwise, and Global Features

Chun Guo
School of Informatics and Computing,
Indiana University Bloomington,
Bloomington, IN, USA, 47405
chunguo@indiana.edu

Yingying Yu
College of Transportation Management,
Dalian Maritime University,
Dalian, China, 116026
uee870927@126.com

Azade Sanjari
School of Informatics and Computing,
Indiana University Bloomington
Bloomington, IN, USA, 47405
asanjari@indiana.edu

Xiaozhong Liu
School of Informatics and Computing,
Indiana University Bloomington
Bloomington, IN, USA, 47405
liu237@indiana.edu

ABSTRACT
The citation relationship between scientific publications has been successfully used for bibliometrics, information retrieval and data mining tasks, and citation-based recommendation algorithms are well documented. While previous studies investigated citation relationships from various viewpoints, most of them share the same assumption that, if $paper_1$ cites $paper_2$ (or $author_1$ cites $author_2$), they are connected, regardless of citation importance, sentiment, reason, topic, or motivation. However, this assumption is oversimplified. In this study, we propose a novel method to automatically label the massive citations in the scientific repository with different roles, a.k.a. citation role labeling. Unlike earlier studies, we employ pairwise features (similarity between citing and cited paper) and global features (citing and cited paper proximity on the heterogeneous graph), in addition to local features (information extracted solely from the citing paper, e.g. citation textual context). Evaluation result shows pairwise and global features, if properly used, can be very helpful to enhance the citation role labeling performance, especially when full-text data is not readily available.

Keywords
Citation Classification, Citation Role Labeling, Machine Learning, Heterogeneous Graph Mining.

77th ASIS&T Annual Meeting, October 31- November 5, 2014, Seattle, WA, USA.

INTRODUCTION
In citation network analysis, the complex citation behavior is reduced to a simple edge, namely, $paper_1$ cites $paper_2$. The implicit assumption is that $paper_1$ is giving credit to, or acknowledging, $paper_2$ (Ding et al., 2013). It is also the case that the contributions of all citations are treated equally, even though the importance, functions and roles of different citations vary from the citing paper point of view. Over the years, researchers have been questioning the validity of this assumption and challenging the usefulness of the oversimplified statistical citation relationships (Bornmann & Daniel, 2008). One critical concern is that when citations are counted, each of them is treated without distinction. However, in reality, citations do not make the same kind of contribution to the citing article, and should not get equal credits. For instance, Shotton (2009) found most existing methods and applications treat all the citations equally, regardless of importance, sentiment, reason, topic, or motivation of the citations, which are potentially important for bibliometrics analysis and scientific data/text mining.

There have been studies that tried to statistically differentiate citations based on their topical contribution. Liu et al. (2013), for instance, inferred the topical importance distribution of citations from full text data using the supervised topic modeling algorithm. Then the PageRank with prior algorithm was able to identify articles that do not have a high citation count but make better contribution to a specific topic. However, for this kind of studies, the roles of citations are still the same, while the topical importance of citations vary.

More recently, researchers started to use text mining and supervised machine learning to automatically categorize citations into different types of functions. In other words,

cited papers make different kinds of contributions to the citing paper. A few studies on automatic identification of citation functions (Athar, 2011; Jochim & Schütze, 2012; Teufel, Siddharthan, & Tidhar, 2006) have been done on Computational Linguistics literature. Text mining and natural language processing techniques were employed to analyze the text surrounding citations in the citing paper. However, there are two major limitations for this approach. First, it is very labor intensive to generate the training data and all the test papers in the collection should have full-text (for citation context extraction). Second, most of these studies extracted local features from citation context (in the citing paper), for instance, bag-of-words, main verb, and part-of-speech features, but the pairwise or global features, i.e., the absolute or relative relationship between citing and cited paper pair, were ignored.

The contribution of this paper is twofold. First, we propose new categorized features for the citation role labeling task. To the matter of greater details, we propose three groups of features in this paper:

1. Local features, extracted from the citation context (in the citing paper), represent the semantic content of the citation function. Local features need full-text corpus.

2. Pairwise features, extracted from the pair of citing and cited papers, characterize the absolute relatedness between the paper pair, i.e., title and abstract similarity of citing and cited papers.

3. Global features, extracted from a large heterogeneous scientific network, characterize the relative relatedness between citing and cited papers with respect to the whole scientific repository. The citing and cited papers are conceptualized as the vertices on the graph, and various types of paths between them, e.g., the path via author, keyword or venue vertex, depict their relationships.

To the best of our knowledge, both pairwise and global features are rarely addressed in previous studies. Moreover, unlike local features, both pairwise and global features are NOT dependent on full-text data, which means we can implement the citation role labeling task with only scientific metadata.

Second, we propose a novel citation role scheme, which is tailored for computer science literature. For instance, we classify all the citations into "Research Question", "Methodology", "Dataset" and "Evaluation", which can be important for computer science paper recommendation and retrieval systems and algorithms.

RELATED WORK

A scholarly article could be cited for different reasons. For this reason, there have been questions about whether citation counts could be used merely as the measures for research impact or acknowledgement. As Cronin has put it, there haven't been any explicitly formalized conventions about citation practice, and thus, it is not clear what is being measured by citation counts (Cronin, 1982).

Most recent studies in text mining, bibliometrics, and scholar information retrieval/recommendation used statistical "citation weighting" approach to differentiate citations. For instance, CiteRank (Walker et al., 2007) is an enhanced ranking algorithm over PageRank, which enables ranking method to estimate the traffic $T_i(\tau_{dir}, \alpha)$ to a given paper i. For this method, a *recent* citation is more likely to be selected with a probability that is exponentially discounted according to the age of the paper, τ_{dir}. Citation Influence Model (Dietz et al., 2007) is another method to weight the importance of a citation, which employed citing and cited paper topic distribution, and the compatibility-based citation weighting of two topic mixtures is measured by the Jensen-Shannon Divergence. A similar research is implemented by Erosheva et al. (2004), which captures the notion of topical similarity between the contents of the cited and citing documents. Based on these work, Nallapati et al. (2008) proposed Pairwise-Link-LDA and Link-PLSA-LDA, which aim to predict important unseen citation relationship between papers by using topic based graph models. Vice versa, citation relation can also be used to characterize the topic models. However, all of those studies share the classical assumption that all the citations make the same type of contributions to the citing paper, while their importance could be statistically different.

There have been a number of empirical studies that looked

Author(s)	Year	Domain	Dataset Size	Scheme	Main features proposed
Teufel et al.	2006	Computational linguistics	116 papers, 2829 citations	Adapted from Spiegel-Rösing (1977), 21 mutually exclusive categories grouped into 4 top-level categories	Cue phrases, verbal tense, verbal voice, modal verb, location features, self-citation.
Athar	2011	Computational linguistics	310 papers, 8736 citations	Negative/Positive/Objective	N-grams, part-of-speech tags, dependency, polarity lexicon.
Jochim & Schütze	2012	Computational linguistics	84 papers, 2008 citations	Moravcsik & Murugesan (1975)	Named entities, constituent-based features, existence of specific part-of-speech tag.

Table 1: Brief Summary of Existing Automatic Citation Classification Research

into the functions (roles) of citations by means of citation context analysis (Moravcsik & Murugesan, 1975; Spiegel-Rösing, 1977). Researchers attempted to illuminate the functions of citations by devising a classification scheme based on an analysis of the text surrounding the citations. Moravcsik and Murugesan (1975) performed the first in-depth citation context analysis, trying to find out the organic nature of citation counts. They analyzed 30 articles from the field of theoretical high energy phisics and manually judged each citation from four aspects, corresponding to the aim, type of connectedness, contribution, and quality of a citation. Spiegel-Rösing (1977) studied the use of citations in 66 articles published in the journal *Science Studies* during its first 4 years of apppearance. The classification scheme has 13 categories, covering many aspects of the use of citations. Unlike Moravcsik and Murugesan, she didn't extract the orthogonal facets in citation functions, but rather provided a flat list of functions instead. These studies provide different perspectives on interpreting the merit and function of a citation (e.g. whether it is truly needed in the citing article; whether it is erroneous; etc.) However, it is hard to incorporate these perspectives into the current way of counting citations, since the analysis requires the citing paper to be read to make proper judgments about each citation and, thus, cannot be easily applied to a large scientific digital library.

Recently, the task of automatic citation classification has caught the attention of computational linguists. Teufel, Siddharthan, and Tidhar (2006), for example, developed a citation classification scheme based on Spiegel-Rösing's (Spiegel-Rösing, 1977). Their scheme contains 12 categories that are mutually exclusive. Their experimental data are made up of 116 conference articles in computational linguistics. They employed cue phrases, verbal tense & voice, modal verb, citation location, and self-citation as features in the classification model. Later, Athar (2011) applied sentiment analysis to citation classification. He manually annotated 8,736 citations from 310 papers from the ACL Anthology into three categories: positive, negative and objective. His results demonstrate that using trigrams and dependencies can optimize the performance of citation sentiment analysis. In the most recent study on this topic, Jochim and Schütze (2012) directly adopted the faceted citation classification scheme of Moravcsik and Murugesan's (1975) without modification. As mentioned earlier, this scheme consists of four facets, conceptual vs. operational, organic vs. perfunctory, evolutionary vs. juxtapositional, and confirmative vs. negational. Their dataset consists of 84 documents from the ACL 2004 proceedings. They used a combination of features from previous studies and proposed on their own and evaluated the performance of different feature groups. More detailed comparison of existing studies is listed in Table 1. As we can see, the datasets used in these studies all come from the computational linguistics domain, mainly for two reasons. First, annotating citation contexts requires in-depth domain knowledge, which limits the subject coverage of experimental data. Second, the features proposed in these studies rely on full-text data to implement and such kind of dataset is available in the computational linguistics domain. However, for most scholarly repositories, full-text is not readily available, which makes it difficult to perform automatic citation role labeling. In this study, we propose new features for this task, which only require scientific metadata to implement.

METHODOLOGY

In this section, we propose the detailed method for citation role labeling. We first discuss the citation role labels used in this study, and then propose categorized features for automatic citation role labeling.

Citation Labels

As mentioned above, a number of studies motivate us to differentiate citation roles and functions, which provides theoretical foundation for this work. These studies characterize and categorize citation roles from different perspectives. Two criteria were involved when we select our annotation scheme: 1. The scheme should be potentially useful for scientific retrieval and recommendation systems. 2. It should be feasible (and not too complicated) for human annotators to make reliable judgments. In other words, for the manual annotating process, a reasonable agreement rate should be achieved between the human annotators, which is a prerequisite for supervised machine learning tasks.

In this research, we adapted the faceted classification scheme originally proposed by Moravcsik and Murugesan (1975), which is also the scheme used by Jochim and Schutze (2012) in their automatic citation classification task. We agree with Jochim and Schutze that making several binary decisions is easier than making a single decision that involves picking the best fit from more than 10 categories. So a faceted classification scheme is more plausible than a flat list. However, unlike Jochim and Schutze, we did not directly adopt Moravcsik and Murugesan's scheme. A few adaptations were made so the resulting categories fit better to papers in the domain that we are interested in (computer science) and could be practical for future development of information retrieval and recommendation applications.

Moravcsik & Murugesan initially proposed their scheme in 1975 and then provided more details about the scheme in a following-up paper in 1978 (Murugesan & Moravcsik, 1978). The citation role scheme has four different facets: Conceptual vs. Operational, Organic vs. Perfunctory, Evolutional vs. Juxtapositional, and Confirmative vs. Negational. We discarded the Confirmative vs. Negational facet in our experiment, because prior studies have shown that authors seldom use negative citations (Bornmann & Daniel, 2008). In the next paragraphs, we elaborate on the other three facets and our adaptations to them.

Research Question vs. Methodology vs. Dataset vs. Evaluation

The Research Question vs. Methodology vs. Dataset vs. Evaluation facet crystallizes to the conceptual vs. operational facet in Moravcsik & Murugesan's original scheme, which distinguishes concept/theory from tools/mathematical techniques. In our research we further split the operational category to three finer ones. The rationale behind this design lies in the nature of computer science research. To a large extent, computer science is an empirical science, which requires formal procedures to guarantee the validity of research results. These formal procedures require standard experiment datasets and reasonable evaluation metrics. Besides, the Research Question vs. Methodology vs. Dataset vs. Evaluation structure has also been employed to classify scientific keywords for information retrieval purposes (Guo, Chinchankar, & Liu, 2012).

Organic vs. Perfunctory

This facet distinguishes whether a citation is "really necessary for the development of the citing paper" (p142, Murugesan & Moravcsik, 1978). For example, the authors might have surveyed several algorithms that would solve the problem, but finally picked the most suitable one. While they might have cited all algorithms, only the adopted one should be annotated as organic.

Evolutionary vs. Juxtapositional

The goal of this facet is to "distinguish material in the same line of work from material in parallel or divergent lines" (p143, Murugesan & Moravcsik, 1978). When researchers write a scientific paper, it is a common practice to explain how their work connects to existing ones, especially those within the same line of research. This facet of the classification scheme helps build connections between papers on the same research topic. When we visualize these connections on the whole citation graph, we would be able to see the evolution of research topics from the bird view.

Features

Earlier studies mainly rely on the textual or linguistic features extracted from the citing article itself, especially the text surrounding citations. In this study, we explore additional novel features that build on the explicit and implicit relationships between the citing and cited article. We group features into three categories: local, pairwise, and global.

Local features

Most features in this group inherit from existing studies. In the most recent work on automatic citation classification, Jochim and Schütze (2012) reimplemented most features from previous studies and evaluated the effectiveness of different feature groups. In this study, we merely implemented the ones that were proved to be important in their experiment, including unigrams, main verb, presence of specific word part-of-speech tags, citation location, etc. Their detailed descriptions can be found in Table 2.

So far, most prior studies in this area were performed by computational linguists, which explains why most local features are linguistically based. For example, the main verb of a sentence can be located by dependency parsing, which is usually directly dependent on the sentence root. This feature is motivated by the fact that authors tend to use specific verbs when making certain types of citations. For example, when citing a paper to clarify the research question, authors tend to use verbs like "explore" or "investigate", instead of the verbs like "use" and "apply", which are more often used to refer to a method.

However, there are also features from this group that do not rely on natural language processing techniques, e.g. location features. For example, intuitively, citations motivated by the research question are more likely to appear at the beginning of the citing article.

Note that, all the local features share two characters in common. First, all the features are extracted merely from the citing paper. Second, full-text citing paper is required for local feature extraction.

Pairwise features

Local features simply reflect the role of a citation in terms of rhetoric style, which is topically independent. In this study, we argue that topic relatedness, between citing and cited paper, is also an important indicator in distinguishing between different citation roles. For example, an evolutionary citation would be topically more related to the citing article than a juxtapositional one.

In this study, we represent topic relatedness in terms of textual similarity. Title similarity and abstract similarity were calculated for each citing-cited article pair. Since the calculation only involves the citing and cited paper, we group them as pairwise features. For pairwise features, textual scientific metadata, i.e., the titles and abstracts of citing and cited papers, are used.

Global features

Global features characterize citing and cited paper relatedness from the viewpoint of the scientific global heterogeneous network, i.e., the citing and cited paper relationship via author, venue and citation random walk on the pre-defined meta-paths. For instance, some evolutionary citation relations might connect citing and cited papers via the same scientific keyword(s) or the same venue.

Unlike local and pairwise features, which mine informative knowledge via citation context and citing-cited paper pair, global features harvest citing and cited paper "relative relatedness" with respect to the whole scientific repository. In other words, citing and cited papers are positioned (as vertices) in a global heterogeneous scholarly graph, and the relationship between them is characterized by a number of paths on the graph between the vertex pair. Note that,

Feature Group	Name	Type	Description
Local Features	Unigram	Boolean	Unigrams
	Voice	Boolean	Verbal voice
	Modal	Nominal	Modal verb (if any)
	Has-modal	Boolean	Sentence has modal verb
	Root	Nominal	Dependency root node
	Main verb	Nominal	Main verb of the sentence
	Has-1stPRP	Boolean	Has first person POS
	Has-3rdPRP	Boolean	Has third person POS
	Comp/sup	Boolean	Has comparative/superlative POS
	But	Boolean	Has "but"
	Has-cf	Boolean	Has "cf."
	PaperLoc	Nominal	Location in the first quarter, middle half (25-75%), or last quarter of the paper.
	SecLoc	Nominal	Location in the first quarter, middle half (25-75%), or last quarter of the section.
	ParaLoc	Nominal	Location in the first quarter, middle half (25-75%), or last quarter of the paragraph.
	SentLoc	Nominal	Location in the first quarter, middle half (25-75%), or last quarter of the sentence.
Pairwise Features	Title-similarity	Real	Title similarity
	Abstract-similarity	Real	Abstract similarity
Global Features	$P_{citing} \rightarrow A \rightarrow A \leftarrow P_{cited}$	Real	The authors of both citing paper and cited paper have collaborated with each other.
	$P_{citing} \rightarrow V \leftarrow K \rightarrow V \leftarrow P_{cited}$	Real	The venues, where the citing paper and cited paper published, contribute to the same topics.
	$P_{citing} \rightarrow P \leftarrow P_{cited}$	Real	Citing paper and cited paper cite the same papers.
	$P_{citing} \rightarrow K \leftarrow P_{cited}$ (relevant edge)	Real	Citing paper and cited paper are relevant to the same topics.
	$P_{citing} \rightarrow K \leftarrow P_{cited}$ (contribution edge)	Real	Citing paper and cited paper contribute to the same topics.

Table 2: Feature List

various kinds of paths may exist on the graph to depict the relationship. For instance, from paper author perspective, $P_{citing} \rightarrow A \rightarrow A \leftarrow P_{cited}$ tells the citing and cited paper relationship via authorship and co-author linkage. A more detailed graphical feature list is presented in Table 2, where the candidate paths and path descriptions are provided.

For this study, a heterogeneous graph with four different types of vertices, *paper, author, keyword,* and *venue*, is constructed. On the graph, five different types of edges connect different vertices:

1. $P \rightarrow A$: paper is written by an author;
2. $P \rightarrow V$: paper is published in a venue;
3. $P \rightarrow K$ *(Relevant)*: paper is relevant to a keyword;
4. $K \rightarrow P$ *(Contribution)*: keyword (topic) is contributed by a paper; and
5. $K \rightarrow V$: keyword (topic) is contributed by a venue.

Edges 1, 2, and 3 are implemented by using scholarly metadata. Paper and venue's contribution to keyword labeled topic, edge 4, and 5, is calculated using the PageRank with Prior algorithm (Liu, Zhang, and Guo, 2013) on the homogeneous citation graphs (paper citation graph and venue citation graph).

$$\pi_{key_t}(v)^{i+1} = \left(1-\beta_b\right)\left(\sum_{u=1}^{d_{in}(v)} p(v|u)\,\pi_{key_t}{}^{i+1}(u)\right) + \beta_b p_{key_t}(v)$$

, where $\pi_{key_t}{}^{i+1}(v)$ is the vertex v's (paper or venue) authority vector over each candidate keyword key_t, and $p_{key_t}(v)$ is the vertex prior distribution (the probability that paper or venue is relevant to the target keyword $\mathrm{key_t}$). The transitioning probability $p(v|u)$, from vertex u to vertex v, is calculated by using the citation relationship (probability) between any pair of vertices on the homogeneous graph. Finally, the authority score $\pi_{key_t}{}^{i+1}(v)$ is used as the contribution edge's weight (for edge type 4 and 5) on the heterogeneous graph. Given space limitation, more detailed description of the PageRank with Prior algorithm can be found in (Liu, Zhang, and Guo, 2013).

For each global feature, we use the random walk probability of the respective meta-path as the feature value. Following each path $\mathcal{P}$, the random walk probability is calculated by:

$$RW_{\mathcal{P}}\left(P_{citing} \rightsquigarrow P_{cited}\right) = \prod_{v_i \in \mathcal{P}} P(v_{i+1}|v_i)$$

, where v_i is a vertex on the path $\mathcal{P}$, and the random walk is defined by multiplying all the transitioning probabilities between each vertex pair on the path. More details about heterogeneous scholar network construction and mining could be found in (Liu, Yu, Guo, Sun, and Gao, 2014).

While the usefulness of the global features is awaiting the feature performance evaluation, compared with local and pairwise features, global features have several advantages. First, global features are more heterogeneous than local and pairwise features, and different paths depict the citing and cited relationship from different perspectives. Second, global features are less sensitive to the quality of scientific repository. For instance, local features require the acquisition of full-text data (for citation context mining) and pairwise features need high quality publication title and abstract (textual metadata), which may not be readily available. Global features, on the contrary, only depend on the scientific metadata, i.e., paper authorship, citation relationship, paper-keyword relationship, and venue information. When full-text information is not available, global features play an important role in the automatic citation role labeling model.

EXPERIMENT

In this section, we describe the experimental setting and results. Analysis and discussion are presented in the next section.

Data

Citation role labeling is a supervised learning task and its performance relies on a high quality annotated dataset. Our dataset is made up of 54 full papers from the Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011). Each in-text citation from these 54 papers is manually labeled three times for the three citation role facets. A paper might be cited more than once in the citing article. In such cases, it is possible for the same paper to perform different roles each time it appears in the citing article. Two graduate students with information retrieval backgrounds were hired to annotate the dataset. Before they started working, we trained the annotators for six weeks in face-to-face meetings and with coding practice. After training, the coders achieved an agreement of 70% or above for all facets. A different dataset were used for annotator training (randomly selected SIGIR full papers published before 2011). In the actual coding phase, each annotator was assigned to work on 30 papers independently. Since correct annotation highly rely on the understanding of the papers, they can skip a paper if they feel they are lacking of background knowledge to make reliable judgments. As mentioned above, the final annotated dataset contains 54 papers, from which, we extracted 2,156 annotated citation contexts. The 2,156 instances were used for machine learning model training and for evaluation purposes.

Full text of the 54 SIGIR papers is converted from PDF documents using the method described in (Gao et al., 2011). During the conversion, original structure of the document (e.g. sections, paragraphs) is roughly retained. Some manual work was involved afterwards to clean up any mistakes caused during the conversion. Section and paragraph structures are important to citation role labeling

task, since citation locations within paragraph and section are important local features.

One motivation of our research is to find automatic citation labeling features that do not solely rely on full-text data. In addition to the full-text data from the 54 SIGIR papers, we also used metadata from two scholarly databases (ACM digital library and CiteSeer) for pairwise and global feature extraction. ACM database has complete metadata for the 54 annotated papers, but does not have complete metadata for the papers that they cited. However, titles of the cited papers are available in the ACM database. Using these titles, we could recover missing metadata for these papers from CiteSeer, which is a much larger scholarly database, through title match.

Citation Context Identification

We extracted citation contexts from the 54 annotated papers using regular expression. SIGIR proceedings follow ACM format and the in-text citations are number(s) wrapped in square brackets (numbers are separated by commas if multiple papers are cited together). We consider each square bracket as an individual citation context. Multiple papers cited in the same bracket (e.g. [1, 3, 5, 12]) are treated as one single citation context. Thus, papers within the same citation context share the same citation role.

Feature Extraction

Local feature

In this study, we only used the sentence that contains the target citation(s) for linguistic feature extraction. Its neighboring sentences were not taken into account.

Pairwise feature

The similarity between title pairs and abstract pairs are measured by cosine similarity (Salton & McGill, 1983). In cases where a context contains multiple citations, we took the average of similarity measures calculated from individual citations.

Global feature

For all the papers in the ACM digital library, we constructed a heterogeneous graph based on the method presented in methodology section. The graph has 641,930 nodes and 4,702,480 edges. More statistics about the graph can be found in Table 3.

Node/Edge Type	Number
Paper node (P)	249,379
Author node (A)	387,932
Topic node (K)	3,910
Venue node (V)	709
$P \rightarrow A$ (writtenby)	632,074
$A \rightarrow A$ (co-author)	1,379,240
$P \rightarrow K$ (relevant)	804,074
$P \leftarrow K$ (contribute)	660,037
$P \rightarrow V$ (publishedat)	247,030
$P \rightarrow P$ (cite)	755,162
$V \leftarrow K$ (contribute)	224,863

Table 3: Typed Nodes and Edges on the Heterogeneous Graph

Machine Learning Algorithm

We used Random Forest (Breiman, 2001), a state of the art machine learning algorithm, for this classification task. It is an ensemble learning method. The basic idea is to construct multiple trees with randomly selected subsets of features during training time. When it comes to predicting, the classifier output the class that is the mode of the classes predicted by individual trees.

Evaluation

In order to validate the usefulness of pairwise and global features for the citation role labeling task, we implement experiments with different feature group combinations:

1. Local features (L)

2. Local features + Pairwise features (L+P)

3. Local features + Global features (L+G)

4. Local features + Pairwise features + Global features (L+P+G)

5. Pairwise features + Global features (P+R)

The local feature group is used as the base feature set. Combinations 2-4 are implemented to test whether Pairwise and/or Global features bring performance gain to the classification task. As mentioned above, local features need full-text citing paper content for citation context extraction. Pairwise and global features only utilize scholarly metadata. Feature group 5, as a result, doesn't need full-text data.

Evaluation, in this study, is performed by using 10-fold cross validation. We report accuracy for each experimental setting as well as precision (P), recall (R) and F-measure (F) for individual classes.

As the baseline, for each facet, the setting in which the class with the highest number of instances is used to label all the citation instances. Accuracy in this setting is equal to the percentage of instances that belong to the majority class.

Feature Group	Research Question			Methodology			Dataset			Evaluation			Acc.
	P	R	F	P	R	F	P	R	F	P	R	F	
Baseline	0	0	0	0.64	1.00	0.78	0	0	0	0	0	0	63.64%
L	0.76	0.42	0.54	**0.73**	0.95	**0.83**	**0.87**	0.17	0.29	0.79	**0.32**	**0.46**	73.79%
L+P	**0.80**	**0.44**	**0.57**	**0.73**	**0.96**	**0.83**	0.75	0.16	0.26	0.77	0.29	0.42	**74.30%**
L+G	0.76	**0.44**	0.56	**0.73**	**0.96**	**0.83**	0.81	0.17	0.29	**0.83**	0.29	0.42	74.12%
L+P+G	0.78	0.39	0.52	0.72	**0.96**	0.82	0.79	**0.20**	**0.32**	0.78	0.27	0.41	73.14%
P+G	0.35	0.17	0.23	0.67	0.87	0.76	0.30	0.08	0.13	0.44	0.24	0.31	61.78%

Table 4: Cross-Validation Results for the Research Question vs. Methodology vs. Dataset vs. Evaluation Facet

RESULTS AND DISCUSSION

Research Question vs. Methodology vs. Dataset vs. Evaluation

As Table 4 shows, the combination of Local and Pairwise (L+P) features achieved the highest overall classification accuracy for this facet. When looking into individual classes, we find that the Research Question class benefits most from the addition of Pairwise features. A possible explanation for this might be that citations motivated by Research Question are usually topically related to the citing article. It is also apparent from Table 4 that this classification task highly relies on the local features. When only using the pairwise and global features (P+G), the overall accuracy significantly dropped to even below the baseline.

Organic vs. Perfunctory

Feature Group	Organic			Perfunctory			Acc.
	P	R	F	P	R	F	
Baseline	0	0	0	0.58	1.00	0.73	58.44%
L	**0.70**	**0.61**	**0.65**	**0.75**	**0.82**	**0.78**	**72.91%**
L+P	0.69	0.60	0.64	0.74	0.81	0.77	72.31%
L+G	0.66	0.58	0.62	0.72	0.79	0.76	70.27%
L+P+G	0.68	0.58	0.62	0.73	0.81	0.77	71.15%
P+G	0.48	0.39	0.43	0.62	0.70	0.66	56.91%

Table 5: Cross-Validation Results for the Organic vs. Perfunctory Facet

For this facet, pairwise and global features (P+G) cannot bring any performance gain, and caused the accuracy to drop. The result indicates that whether a citation is truly needed is not closely related to its topical similarity to the citing article, or its heterogeneous connectivity with the citing article in the scholarly network. For example, an author might cite a few papers to acknowledge existing methods for solving similar research questions. If the author proposes a new method for an existing research problem, even though these citations are all topically related, they are perfunctory. On the other hand, if the author were not proposing a new method, but instead, merely optimizing one of these existing methods, the paper for this particular method would be an organic citation, since the citing article is building on top of it. In contrast, local features have an obvious advantage. If the author explicitly uses the words with strong linguistic clue, e.g., "used" "utilized" or "employed", we can highly likely tell that this citation is an organic one. This is a case where main verb in the citation sentence can effectively distinguish an organic citation from the perfunctory ones. Given these reasons, it is not surprising to see that Local features (L) outperformed Pairwise and Global features (P+G) in the evaluation.

Evolutionary vs. Juxtapositional

Feature Group	Evolutionary			Juxtapositional			Acc.
	P	R	F	P	R	F	
Baseline	0	0	0	0.73	1.00	0.84	72.77%
L	0.57	0.31	0.40	0.73	**1.00**	0.84	74.68%
L+P	0.54	0.26	0.35	**0.78**	0.91	0.84	73.79%
L+G	0.59	0.26	0.36	0.77	0.92	0.84	74.81%
L+P+G	0.63	0.25	0.36	0.77	0.93	0.85	75.60%
P+G	**0.81**	**0.53**	**0.64**	0.77	0.95	**0.90**	**83.86%**

Table 6: Cross-Validation Results for the Evolutionary vs. Juxtapositional Facet

For the Evolutionary vs. Juxtapositional facet, we can observe a significant performance gain by using pairwise features and global features. Although adding Pairwise features to the Local feature group (L+P) makes accuracy drop a little, precision for the majority class is improved. We observe that recall for the local feature group (L) is 1, which indicates that it is highly likely the classifier is biased towards the majority class. Adding pairwise features into the model can effectively help the classifier recover the true concept that distinguishes one class from the other. Global features, for this facet, bring in much more benefits, as the combination of global features and local features (L+G) gains better prediction accuracy than using local features

alone. Accuracy is further improved when we appended both Pairwise and Global features to the Local feature group (L+P+G). It is interesting to find that using the Pairwise and Global features alone (P+G) yields the best accuracy in the sets of experiments. Experiment results suggest that local features might be less informative for this classification task, which also distinguishes this facet from the other two: whether a citation is evolutionary or juxtapositional is not merely dependent on the citation context. The similarity between citing-cited paper pair, and the heterogeneous relatedness between the citing and cited articles on the graph are more important for this facet. In other words, the same cited article might be labeled differently with different citation contexts for the other two facets. But for the Evolutionary vs. Juxtapositional facet, it is highly likely that the same cited paper serves the citing article with one unique role.

CONCLUSION

In this paper, we propose innovative categorized features for citation role labeling tasks. Unlike earlier studies focusing on local features extracted from citing full-text papers, we investigate pairwise and global features to characterize the absolute and relative relationships between citing and cited papers. Pairwise features calculate the similarity between citing and cited papers, while global features explore the relative relatedness between citing and cited papers via different kinds of paths on the heterogeneous graph. Pairwise and global features are extracted from scholarly metadata, so that full-text is no longer the premise of the citation role labeling tasks.

Evaluation result shows that:

1. Pairwise and global features (P+G) are very useful for Evolutionary vs. Juxtapositional facet labeling. The pairwise and global features' performance significantly outperforms other types of features. We assume that, as this facet helps estimate the evolutionary connections between papers on the same research topic, the paper similarity and the interim path between citing and cited papers via keyword, author, and venue can be significant features for this facet.

2. For Research Question vs. Methodology vs. Evaluation vs. Dataset facet, merely using pairwise and global features cannot improve classification performance. However, we find when we use pairwise or global features along with the local features (L+P or L+G), the classification performance can be enhanced.

3. Local textual features surpass other kinds of features for Organic vs. Perfunctory facet. This could be the reason that text features provide strongest clue for organic citations, e.g., "based on [citation], we develop…". Compared with local features, pairwise or global features may not be appropriate for this facet. For example, a computer scientist could investigate two different kinds of algorithms, and the path via authorship doesn't necessarily indicate these two studies are organically connected.

In sum, we find pairwise and global features, if properly used, can be very helpful to enhance the citation role labeling performance, especially when full-text data is not readily available.

FUTURE WORK

This study shares a common limitation with existing works on citation classification: the experiment dataset is small and comes from a specific domain. Thus, the generalizability of the results is questionable. But, since the generation of experimental dataset requires a fair amount of human annotation from domain experts, this limitation is hard to avoid.

To further prove the effectiveness of the method proposed in this study, we need to test it on a larger corpus. In the future, we plan to label the citations in a large computer science literature database with the model we've derived. Later on, we can build a retrieval system, where retrieved papers would be clustered based on the roles they have played in other papers as citations. We will evaluate the usefulness and applicability of the labeled citation roles as part of the usefulness of the retrieval system.

REFERENCES

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session* (pp. 81-87). Portland, Oregon: Association for Computational Linguistics.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of documentation, 64*(1), 45-80.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Cronin, B. (1982). Norms and functions in citation: The view of journal editors and referees in psychology. *Social Science Information Studies, 2*(2), 65-77.

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine learning* (pp. 233-240). Corvallis, OR: ACM.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United states of America, 101*(Suppl 1), 5220-5227.

Gao, L., Tang, Z., Lin, X., Liu, Y., Qiu, R., & Wang, Y. (2011). Structure extraction from PDF-based book documents. In *Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 11-20), Ottawa, Ontario, Canada: ACM.

Guo, C., Chinchankar, R., & Liu, X. (2012). Knowledge retrieval for scientific literatures. *Proceedings of the*

American Society for Information Science and Technology, 49(1), 1-7.

Jochim, C., & Schütze, H. (2012). Towards a generic and flexible citation classifier based on a faceted classification scheme, In *Proceedings of the 2012 International Conference on Computational Linguistics* (pp. 1343-1358). Mumbai, India: The COLING 2012 Organizing Committee.

Liu, X., Yu, Y., Guo, C., Sun, Y., & Gao, L. (2014). Context-rich heterogeneous network mining approach for citation recommendation. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL 2014). London, UK: ACM.

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.

Moravcsik, M. J., & Murugesan, P. (1975). Some Results on the Function and Quality of Citations. *Social Studies of Science, 5*(1), 86-92.

Murugesan, P., & Moravcsik, M. J. (1978). Variation of the nature of citation measures with journals and scientific specialties. *Journal of the American Society for Information Science, 29*(3), 141-147.

Nallapati, R. M., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008). Joint latent topic models for text and citations, In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 542-550). Las Vegas, NV: ACM.

Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill, Inc.

Shotton, D., Portwin, K., Klyne, G., & Miles, A. (2009). Adventures in semantic publishing: exemplar semantic enhancements of a research article. *PLoS Computational Biology, 5*(4), e1000361.

Spiegel-Rösing, I. (1977). Science Studies: Bibliometric and Content Analysis. *Social Studies of Science, 7*(1), 97-113.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function, In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 103-110). Sydney, Australia: Association for Computational Linguistics.

Walker, D., Xie, H., Yan, K.-K., & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment, 2007*(6), P06010.