Co-Training for Topic Classification of Scholarly Data

Cornelia Caragea[1], **Florin Bulgarov**[1], **Rada Mihalcea**[2]
[1]Computer Science and Engineering, University of North Texas, TX, USA
[2]Computer Science and Engineering, University of Michigan, MI, USA
ccaragea@unt.edu, FlorinBulgarov@my.unt.edu,mihalcea@umich.edu

Abstract

With the exponential growth of scholarly data during the past few years, effective methods for topic classification are greatly needed. Current approaches usually require large amounts of expensive labeled data in order to make accurate predictions. In this paper, we posit that, in addition to a research article's textual content, its citation network also contains valuable information. We describe a co-training approach that uses the text and citation information of a research article as two different views to predict the topic of an article. We show that this method improves significantly over the individual classifiers, while also bringing a substantial reduction in the amount of labeled data required for training accurate classifiers.

1 Introduction

As science advances, scientists around the world continue to produce a large number of research articles, which provide the technological basis for worldwide dissemination of scientific discoveries. Online digital libraries such as Google Scholar, CiteSeer^{x}, and PubMed store and index millions of such research articles and their metadata, and make it easier for researchers to search for scientific information. These libraries require *effective* and *efficient* methods for topic classification of research articles in order to facilitate the retrieval of content that is tailored to the interests of specific individuals or groups. Supervised approaches for topic classification of research articles have been developed, which generally use either the content of the articles (Caragea et al., 2011), or take into account the citation relation between research articles (Lu and Getoor, 2003).

To be successful, these supervised approaches assume the availability of large amounts of labeled data, which require intensive human labeling effort. In this paper, we explore a semi-supervised approach that can exploit large amounts of unlabeled data together with small amounts of labeled data for accurate topic classification of research articles, while minimizing the human effort required for data labeling. In the scholarly domain, research articles (or papers) are highly interconnected in giant citation networks, in which papers cite or are cited by other papers. We posit that, in addition to a document's textual content and its local neighborhood in the citation network, other information exists that has the potential to improve topic classification. For example, in a citation network, information flows from one paper to another via the citation relation (Shi et al., 2010). This information flow and the topical influence of one paper on another are specifically captured by means of citation contexts, i.e., short text segments surrounding a citation's mention.

These contexts are not arbitrary, but they often serve as brief summaries of a cited paper. We therefore hypothesize that these *micro-summaries* can be successfully used as an independent view of a research article in a co-training framework to reduce the amount of labeled data needed for the task of topic classification.

The idea of using terms from citation contexts stems from the analysis of hyperlinks and the graph structure of the Web, which are instrumental in Web search (Manning et al., 2008). Many search engines follow the intuition that the anchor text pointing to a page is a good descriptor of its content, and thus anchor text terms are used as additional index terms for a target webpage. The use of links and anchor text was thoroughly researched for information retrieval (Koolen and Kamps, 2010), broadening a user's search (Chakrabarti et al., 1998), query refinement (Kraft and Zien, 2004), and enriching document representations (Metzler et al., 2009). Blum and

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2357–2366,
Lisbon, Portugal, 17-21 September 2015.

Mitchell (1998) introduced the co-training algorithm using hyperlinks and anchor text as a second, independent view of the data for classifying webpages, in addition to a webpage content.

Contributions and Organization. We present a co-training approach to topic classification of research papers that effectively incorporates information from a citation network, in addition to the information contained in each paper. The result of this classification task will aid indexing of documents in digital libraries, and hence, will lead to improved organization, search, retrieval, and recommendation of scientific documents. Our contributions are as follows:

- We propose the use of citation contexts as an additional view in a co-training approach, which results in high accuracy classifiers. To our knowledge, this has not been addressed in the literature.
- We show experimentally that our co-training classifiers significantly outperform: (1) supervised classifiers trained using either content or citation contexts independently, for the same fraction of labeled data; and (2) several other semi-supervised classifiers, trained on the same fractions of labeled and unlabeled data as co-training.
- We also show that using the citation context information available in citation networks, the human effort involved in data labeling for training accurate classifiers can be largely reduced. Our co-training classifiers trained on a very small sample of labeled data and a large sample of unlabeled data yield accurate topic classification of research articles.

The rest of the paper is organized as follows. In Section 2, we discuss related work. Section 3 describes our data and its characteristics, followed by the presentation of our proposed co-training approach in Section 4. We present experiments and results in Section 5, and conclude the paper and present future directions of our work in Section 6.

2 Related Work

We discuss here the most relevant works to our study. A large variety of methods have been proposed in the literature with regard to automatic text classification and topic prediction. Different classifiers have been applied on the Vector Space Model (VSM), in which a document is represented as a vector of words or phrases associated with their TF-IDF score, i.e. *term frequency - inverse document frequency* (Zhang et al., 2011; Kansheng et al., 2011). VSM is the most used method due to its simple, efficient and easy to understand implementation. Another widely used model is the Latent Semantic Indexing (LSI) where co-occurrences are analyzed to find semantic relationships between words or phrases (Zhang et al., 2011; Ganiz et al., 2011). Moreover, a great range of classifiers were used for this task, including: Naïve Bayes (Lewis and Ringuette, 1994), K-nearest neighbors (Yang, 1999) and Support Vector Machines (Joachims, 1998). These techniques, however, all require a large number of labeled documents in order to build accurate classifiers. In contrast, we propose a co-training algorithm that only requires a small amount of labeled data in order to make accurate topic classification.

Semi-supervised methods essentially involve different means of transferring labels from *labeled* to *unlabeled* samples in the process of learning a classifier that can generalize well on new unseen data. Co-training was originally introduced in (Blum and Mitchell, 1998) where it was used to classify web pages into *academic course home page* or not. This approach has two views of the data as follows: the content of a web page, and the words found in the anchor text of the hyperlinks that point to the web page. Wan (2009) used co-training for cross-lingual sentiment classification of product reviews, where English and Chinese features were considered as two independent views of the data. Furthermore, Gollapalli et al. (2013) used co-training to identify authors' homepages from the current-day university websites. The paper presents novel features, extracted from the URL of a page, that were used in conjunction with content features, forming two complementary views of the data.

Citation networks have been used before in other problems. Caragea et al. (2014) used citation contexts to extract informative features for keyphrase extraction. Lu and Getoor (2003) proposed an approach for document classification that used only citation links, without any textual data from the citation contexts. Ritchie et al. (2006) used a combination of terms from citation contexts and existing index terms of a paper to improve indexing of cited papers. Citation contexts were also used to improve the performance of citation recommendation systems (Kataria et al., 2010) and to study author influence in document networks

(Kataria et al., 2011). Moreover, citation contexts were used for scientific paper summarization (Abu-Jbara and Radev, 2011; Qazvinian et al., 2010; Qazvinian and Radev, 2008; Mei and Zhai, 2008; Lehnert et al., 1990) For example, in Qazvinian et al. (2010), a set of important keyphrases is extracted first from the citation contexts in which the paper to be summarized is cited by other papers and then the "best" subset of sentences that contain such keyphrases is returned as the summary. Mei and Zhai (2008) used information from citation contexts to determine what sentences of a paper are of high impact (as measured by the influence of a target paper on further studies of similar or related topics). These sentences constitute the impact-based summary of the paper.

Despite the use of citation contexts and anchor text in many information retrieval and natural language processing tasks, to our knowledge, we are the first to propose the incorporation of citation context information available in citation networks in a co-training framework for topic classification of research papers.

3 Data

The dataset used in our experiments is a subset sampled from the CiteSeer^x digital library[1] and labeled by Dr. Lise Getoor's research group at the University of Maryland. This subset was previously used in several studies including (Lu and Getoor, 2003) and (Kataria et al., 2010). The dataset consists of 3186 labeled papers, with each paper being categorized into one of six classes: Agents, Artificial Intelligence (AI), Information Retrieval (IR), Machine Learning (ML), Human-Computer Interaction (HCI) and Databases (DB). For each paper, we acquire the citation contexts directly from CiteSeer^x. A citation context is defined as a window of n words surrounding a citation mention. We differentiate between cited and citing contexts for a paper as follows: let d be a target paper and $\mathcal{C}$ be a citation network such that $d \in \mathcal{C}$. A *cited context* for d is a context in which d is cited by some paper d_i in $\mathcal{C}$. A *citing context* for d is a context in which d is citing some paper d_j in $\mathcal{C}$. If a paper is cited in multiple contexts within another paper, the contexts are aggregated into a single context. For each paper in the dataset, we have at least one cited or one citing context. A summary of the dataset is provided in Table 1.

[1] http://citeseerx.ist.psu.edu/

Number of papers in each class						
Agents	AI	IR	ML	HCI	DB	Total
562	239	641	569	490	685	3186

Avg. Cited Contexts	Avg. Citing Contexts
45.59	20.77

Table 1: Dataset summary.

As expected, we have a higher number of cited contexts than citing contexts. This is due to the page restrictions often imposed to research articles that can limit the number of papers each article can cite. On the other hand, a good research paper can accumulate hundreds of citations, and hence, cited contexts over the years.

Context lengths. In CiteSeer^x, citation contexts have about 50 words on each side of a citation mention. A previous study by Ritchie et al. (2008) shows that a fixed window length of about 100 words around a citation mention is generally effective for information retrieval tasks. For this reason, we use the contexts provided by CiteSeer^x directly. In future, it would be interesting to study more sophisticated approaches to identifying the text that is relevant to a target citation (Abu-Jbara and Radev, 2012; Teufel, 1999) and study the influence of context lengths on our task.

For all experiments, our labeled dataset is split in train, validation and test sets. The validation and test sets have about 200 papers each. We sampled another set of papers from the labeled dataset in order to simulate the existence of unlabeled data, with a fixed size of around 2000 papers. The remaining 786 papers are used as labeled training data. Each experiment was repeated 10 times with 10 different random seeds and the results were averaged.

4 Co-Training for Topic Classification

Blum and Mitchell (1998) proposed the co-training algorithm in the context of webpage classification. In co-training, the idea is that two classifiers trained on two different views of the data teach one another by re-training each classifier on the data enriched with predicted examples that the other classifier is most confident about. In Blum and Mitchell (1998), webpages are represented using two different views: (1) using terms from webpages' content and (2) using terms from the anchor text of hyperlinks pointing to these pages.

Algorithm 1 Co-Training

Input: L, U, 's'
  $L_1 \leftarrow L$, $L_2 \leftarrow L$
  while $U \neq \varnothing$ **do**
    Train classifier C_1 on L_1
    Train classifier C_2 on L_2
    $S \leftarrow \varnothing$
    Move 's' examples from U to S
    $U \leftarrow U \backslash S$
    $S_1,\ S_2 \leftarrow$ *GetMostConfidentExamples*(S, C_1, C_2)
    $L_1 \leftarrow L_1 \cup S_1$, $L_2 \leftarrow L_2 \cup S_2$
    $U \leftarrow U \cup [S \backslash (S_1 \cup S_2)]$
  end while
Ouput: The combined classifier C of C_1 and C_2

In this paper, we study the applicability and extension of the co-training algorithm to the task of topic classification of research papers, which are embedded in large citation networks. Here, in addition to the information contained in a paper itself, *citing* and *cited* papers capture different aspects (e.g., topicality, domain of study, algorithms used) about the target paper (Teufel et al., 2006), with *citation contexts* playing an instrumental role. We conjecture that citation contexts, which act as brief summaries about a cited paper, provide important clues in predicting the topicality of a target paper. These clues give rise to the design of our co-training based model for topic classification of research papers. In our model, we use the content of a paper as one view and the citation contexts as another view of our data. In particular, for the content of a paper, we use its title and abstract as it is commonly used in the literature (Lu and Getoor, 2003); for the citation contexts, we use both the cited and citing contexts, as described in the previous section.

Our co-training procedure is described in Algorithm 1. L and U represent the labeled and unlabeled datasets and contain instances from both views. The fractions of the training set are obtained from the 786 papers by selecting $k\%$ random examples from each class. For a round of co-training, we train classifiers C_1 and C_2 on the two views. Next, s examples are sampled from the unlabeled data into S, and C_1, C_2 are used to obtain predictions for these s examples. The *GetMostConfidentExamples* method is a generic placeholder that stands for a function that determines what examples from S are chosen to be added into training. Finally, at the end of an iteration, the examples left into S are moved back to U, and the algorithm iterates until there are no more unlabeled examples in U. The final classifier C is obtained by combining C_1 and C_2 using the product of their class probability distributions. The class with the highest posterior probability (of the product of the two distributions) is chosen as the predicted class.

Unlike the original co-training algorithm described by Blum and Mitchell (1998), which tackled a binary classification task (*course* vs. *non-course* page classification), we address a multi-class classification problem, where each example (i.e., research paper) is classified into one of six different classes. Moreover, in Blum and Mitchell (1998), the co-training algorithm moves p highest confidence positive examples and n highest confidence negative examples from S to L, where $p : n$ represents the class distribution in the original labeled training set (i.e., if there are 10 positive examples and 90 negative examples in the labeled set L, then $p = 1$ positive and $n = 9$ negative examples are moved to the labeled set at each iteration of co-training). Unlike, this approach that preserves the class distribution of the original labeled training set, we move into L all examples that are classified with a confidence above a certain threshold.

5 Results and Discussion

First, the proposed method is evaluated on the validation set. We first compare it against various supervised and semi-supervised baselines. Next, we report the performance of our co-training algorithm under different scenarios, where either cited or citing contexts are used. We also show the most informative words for each classifier. Finally, with the best parameters obtained on the validation set, we report the precision, recall and F1-score, obtained by each method, on the test set.

In experiments, the sample size 's' from Algorithm 1 is set to 300, i.e. the number of documents sampled from the unlabeled pool at each iteration; the confidence threshold is set to 0.95, i.e. if both classifiers agree on the class label and have a confidence ≥ 0.95, the instance is labeled and moved into the labeled training set. These parameters are estimated on the validation set, but the results are not shown due to space limitation.

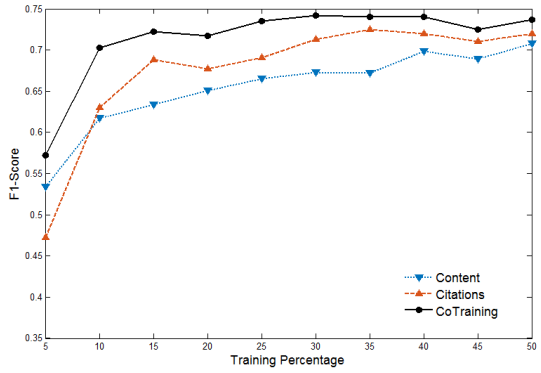

Figure 1: Co-Training vs. Supervised Learning.

Evaluation Measures. We report results averaged over ten different runs with random splits. For each random split, we return the weighted average precision, recall and F1-score. In all the experiments, we use the Naïve Bayes Multinomial classifier and its Weka implementation[2], with term-frequencies as feature values. We experimented with both TF and TF-IDF scores, using different classifiers (Support Vector Machine, Naïve Bayes Multinomial, and simple Naïve Bayes classifiers), but Naive Bayes Multinomial with TF performed best.

5.1 Baseline Comparisons

How does co-training compare with supervised learning techniques? In this experiment, we compare our co-training method with two supervised baselines: (1) when only document content is used and (2) when only citation contexts are used.

Figure 1 shows the F1-scores achieved using different initial training sizes. We can see that overall, the citation contexts are better at predicting the topic of a document compared with the content, outperforming them in 9 out of 10 experimental settings. The only exception to this trend is when a small number (5%) of training instances is available, in which case the supervised content view performs better, reaching an F1-score of 0.534. Regardless, the co-training method shows significant improvement over both baselines, in all experiments. Starting with an F1-score of 0.572, it continues to improve its performance as the training percentage is increasing. The maximum F1-score, i.e. 0.742, is reached when 30% of the labeled training set is used. Note that the difference in performance between co-training and the two supervised baselines is statistically significant for a p value of 0.05.

A fully supervised baseline that uses 100% of the training set achieves an F1-score of 0.720 (using content) and 0.738 (using citation contexts). In contrast, co-training requires only 15% of the labeled training set to outperform the fully supervised content baseline and 30% of the training set to outperform the fully supervised citation contexts baseline. Consequently, using a co-training approach that includes citation contexts as well as the document content can not only increase the performance, but will also significantly reduce the need of expensive labeled instances.

Figure 2 illustrates the confusion matrices of three experiments: (a) supervised content view, i.e. the title and abstract, (b) supervised citation contexts view, and (c) co-training that uses both views. These experiments use 10% of the training set. Each of the matrices are represented by a heat map, i.e. the redder the color, the higher the value assigned to that position. An accuracy of 1 will be represented by a matrix with red blocks on the main diagonal and white blocks everywhere else. This experiment was performed 10 times with 10 different seeds and the results have been averaged.

As can be seen, the matrix that uses only titles and abstracts, i.e. left side, is showing the highest percentage of misclassified documents, classifying correctly about 58.8% instances, on average. Using only citation contexts in a supervised framework, i.e. center matrix, we reach a higher accuracy of 60.7%. The co-training method, which uses the content of the paper and citations as two independent views, significantly increases the average accuracy to 67.3%. This experiment shows that citation contexts are better than titles and abstracts at predicting the topic of a document. Furthermore, our proposed approach, which uses the content of the paper as well as citation contexts, achieves higher results than each view used separately. The difference in accuracy is statistically significant across all three experiments for a p value of 0.05.

Overall, the *Agents* class seem to be the easiest to classify, reaching an accuracy value of 91.6% when using co-training. On the other hand, the *AI* class is the hardest to classify. One reason for this is that the *AI* class contains the lowest number of instances in the dataset. Another can be that the *AI* class is the most general among all classes and therefore, classifying documents with this la-

[2]http://www.cs.waikato.ac.nz/ml/weka/

Figure 2: The accuracy of our method, against two supervised baselines.
Left: using titles and abstracts; Center: using citation contexts; Right: using co-training.

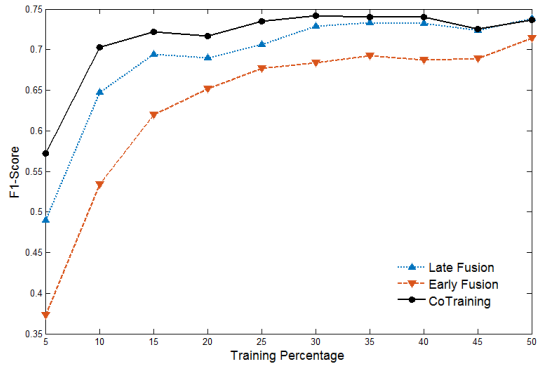

Figure 3: Co-Training vs. Early and Late Fusion.

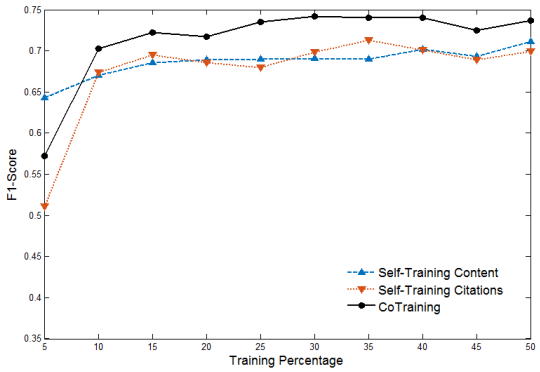

Figure 4: Co-Training vs. Self-Training.

bel can be a difficult task even for a human. Other common misclassifications occur between classes like *HCI* and *Agents*, *ML* and *IR* or *AI* and *ML*, due to their similarity.

How does our co-training method compare with other supervised approaches? In this experiment, we compare the performance of co-training against two other methods: early and late fusion. In early fusion, the feature vectors of the two views are concatenated, creating a single representation of the data. In contrast, late fusion trains two separate classifiers and then combines them by taking the label with the highest confidence.

Figure 3 shows this comparison over different training sizes. The results show that the co-training method is more accurate than all others, performing best in all 10 experimental settings. Late fusion has an overall lower performance compared with co-training, but is in a tight correlation with it. On the other hand, early fusion achieves the lowest F1-score across the experiments. The reported results are statistically significant at p value of 0.05, when the training percentage is between 5 and 35. Therefore, we can say that training two separate classifiers, one of each view, yields higher performance compared with training a single classifier that incorporates both views. Moreover, using a co-training approach that incorporates information from unlabeled data into the model, will help the two classifiers increase their confidences and minimize the error rate.

How does co-training compare with semi-supervised methods? Here, we present results comparing co-training with two other well-known semi-supervised techniques: self-training and Naïve Bayes with Expectation Maximization.

Self-Training. First, we show results of the comparison of co-training with two variations of self-training: (1) self-training using only document content, and (2) self-training using only citation contexts. Figure 4 shows the results of this experiment. Self-training is similar to co-training, except that it uses only one view of the data (Zhu, 2005). Self-training parameters, e.g., sample size 's' or number of iterations, are estimated as in co-training.

Although the document content version of self-training outperforms co-training when using 5%

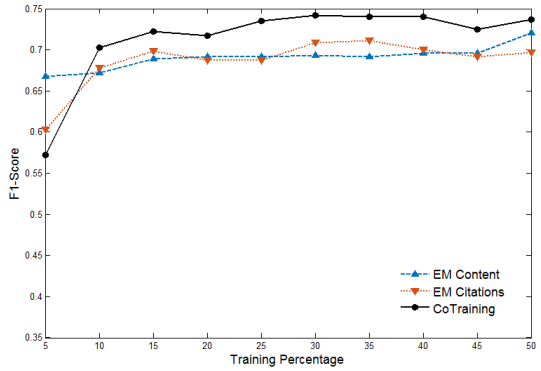

Figure 5: Co-Training vs. EM.

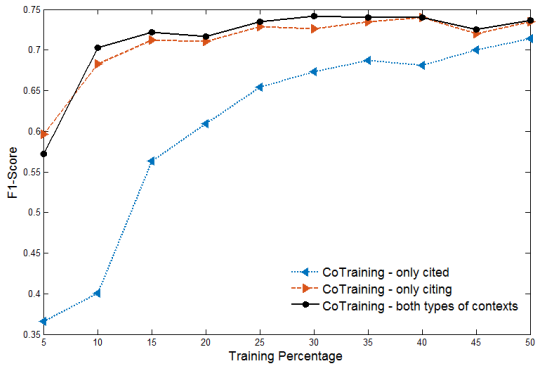

Figure 6: Performance when using only cited / only citing / or both citation contexts.

of the training instances, we can see that overall, there is a significant difference in terms of F1-score values in the favor of co-training. In 9 out of 10 experiments, our co-training approach is superior to both self-training methods. The results are statistically significant across all experimental setups for a p value of 0.05.

Expectation Maximization. Figure 5 shows the F1-score values obtained after running NBM with EM with the same training, unlabeled and test sets. The EM algorithm uses the same classifier, i.e. NBM, and the weight for each unlabeled instance is set to 1, as this setting achieved the highest results. Two different experiments were performed using EM: (1) using only document content, and (2) using only citation contexts. As can be seen in the figure, overall, the co-training approach significantly outperforms both variations of EM. However, the co-training method falls short when using 5% of the training instances, where *EM Content* and *EM Citations* methods are achieving higher F1-score values. Nonetheless, both EM variations tend to achieve an F1-score value below or equal to 0.710, whereas co-training reaches performance values of 0.74 or higher. Again, the comparison results between co-training and both variations of EM are statistically significant for training sizes between 10% and 50%, for a p value of 0.05.

5.2 Using Different Citation Context Types

Which of the two types of citation contexts (cited or citing) help the task of topic classification more and how does co-training perform in the absence of either one? The answer to this question is important as there are cases in which citation contexts are not readily available. One frequently encountered example includes newly published research papers that have no cited contexts. In this case, it is important to know how our method performs when we only have one type of citation contexts. Figure 6 shows the difference in performance when using: (1) only cited contexts, (2) only citing contexts, and (3) both context types. Note that the content view remains the same across all three experiments.

The plot is showing that citing contexts are bringing in a significantly higher margin of knowledge compared with cited contexts. This is consistent over different training set sizes, as shown in the figure, with a more prominent impact when a small training size is used, i.e. 5-30%. The fact that the citing contexts achieve higher F1-score than cited contexts is consistent with the intuition that when citing a paper y, an author generally summarizes the main ideas from y using important words from a target paper x, making the citing contexts to have higher overlap with words from x. In turn, a paper z that cites x may use paraphrasing to summarize ideas from x with words more similar to those from the content of z.

When the two types of contexts are used, co-training achieves higher results compared with cases when only one context type is used. This experiment shows that our method can be applied for both old and new research articles. Citing contexts will be available in the text of the target paper and are independent of the existence of the cited contexts.

5.3 Informative Features

What are the most informative words from each view: document content and citation contexts? Figure 7 shows the words from each view that are most useful for our topic classification task. The larger the word, the more informative is for our

Figure 7: Most informative words from document content (left) and citation contexts (right).

Method	Labeled docs. (%)	Precision	Recall	F1-Score
Co-Training	**30**	**0.749**	**0.743**	**0.742**
Co-Training - only citing	40	0.747	0.740	0.740
Co-Training - only cited	50	0.724	0.717	0.714
Self-Training - Content	50	0.723	0.711	0.711
Self-Training - Citations	35	0.730	0.710	0.713
EM - Content	50	0.738	0.714	0.721
EM - Citations	35	0.729	0.707	0.711
Early Fusion	50	0.718	0.710	0.714
Late Fusion	50	0.748	0.734	0.738
Content - Fully Supervised	100	0.730	0.728	0.720
Citations - Fully Supervised	100	0.745	0.740	0.738

Table 2: A comparison of all methods on the test set.

task. To determine the informativeness of a word, we used its Information Gain score. For these experiments, we used training sets consisting of 30% of the instances, setting in which we achieved the best results on the validation and test sets using our proposed co-training approach.

As can be seen, the two word clouds have a high word overlap. Words such as *agent*, *database* or *query* are almost equally important in the two views, dominating both clouds. However, differences can be observed. For example, words like *learning*, *multi-agent* or *interface* are more important in the content view. On the other hand, words such as *document* or *text* achieve a higher information gain score for the citation contexts view.

5.4 Co-Training vs. All Other Approaches

Table 2 summarizes the results obtained by all the baselines used so far, in comparison with our proposed co-training method. For this experiment, we show the training percentage used, the precision, recall and F1-score for each method, in the setting in which it returned the best results. All measures were averaged after 10 runs with 10 different seeds.

The results in Table 2 show that the proposed co-training method outperforms all compared models, reaching the highest F1-score of 0.742, while using the smallest amount of labeled documents, i.e. 30%. Using only the citing contexts, the performance is similar to that of co-training when both context types are used. However, using only the cited contexts, the performance decreases compared to that of the full model that uses both context types. We see that the citing contexts perform better, reaching an F1-score value of 0.740 compared against 0.714 when only cited contexts are used. Moreover, the method that uses only the citing contexts is using 10% less labeled data.

Self-training and EM show decreased performance compared with co-training. Late Fusion outperforms Early Fusion, i.e., 0.738 vs. 0.714, both obtaining lower results than co-training, while using significantly more labeled data.

The last two lines of the table show the results when all documents (except those in the validation and test), are used for training, in a supervised framework. As can be seen, a supervised method that uses only citations will achieve a higher performance, compared against a method that uses titles and abstracts. Nonetheless, co-training obtains higher results than both fully supervised approaches, while using only 30% of the labeled data.

6 Conclusion and Future Work

In this paper, we studied the problem of using citation contexts in order to predict more accurately the topic of a research article. We showed that a co-training technique, which uses the paper content and its citation contexts as two *conditionally independent* and *sufficient* views of the data, can effectively incorporate cheap, unlabeled data to improve the classification performance and to reduce the need of labeled examples to only a fraction. The results of the experiments showed that the proposed approach performs better than other semi-supervised and supervised methods.

This study also shows that citation contexts are rich sources of information that can be successfully used in various IR and NLP tasks. We showed that document content and citation contexts unified under the same algorithm can dramatically decrease the annotation costs as well. In the future, we plan to extend co-training to include active learning for more robust classification. Moreover, it would be interesting to extend the co-training approach to multi-views that could potentially handle more than two feature spaces, e.g., it could include topics by Latent Dirichlet Allocation (Blei et al., 2003) as an additional view.

Acknowledgments

We are thankful to Dr. Lise Getoor for making the Citeseer^x labeled subset publicly available. We are also grateful to Dr. C. Lee Giles for the CiteSeer^x data, which helped extract the citation contexts of the research papers in the collection. We very much thank our anonymous reviewers for their constructive feedback. This research is supported in part by the NSF award #1423337 to Cornelia Caragea. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF.

References

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 500–509.

Amjad Abu-Jbara and Dragomir Radev. 2012. Reference scope identification in citing sentences. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 80–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100, New York, NY, USA. ACM.

Cornelia Caragea, Adrian Silvescu, Saurabh Kataria, Doina Caragea, and Prasenjit Mitra. 2011. Classifying scientific publications using abstract features. In *The Symposium on Abstraction, Reformulation, and Approximation (SARA)*.

Cornelia Caragea, Adrian Florin Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. 2014. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1435–1446. Association for Computational Linguistics.

Soumen Chakrabarti, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, and Jon Kleinberg. 1998. Automatic resource compilation by analyzing hyperlink structure and associated text. *Comput. Netw. ISDN Syst.*, 30(1-7):65–74, April.

Murat Can Ganiz, Cibin George, and William M Pottenger. 2011. Higher order naive bayes: a novel non-iid approach to text classification. *Knowledge and Data Engineering, IEEE Transactions on*, 23(7):1022–1034.

Sujatha Das Gollapalli, Cornelia Caragea, Prasenjit Mitra, and C. Lee Giles. 2013. Researcher homepage classification using unlabeled data. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 471–482, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Thorsten Joachims. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 137–142, London, UK, UK. Springer-Verlag.

SHI Kansheng, HE Jie, Hai-tao LIU, Nai-tong ZHANG, and Wen-tao SONG. 2011. Efficient text classification method based on improved term reduction and term weighting. *The Journal of China Universities of Posts and Telecommunications*, 18:131–135.

Saurabh Kataria, Prasenjit Mitra, and Sumit Bhatia. 2010. Utilizing context in generative bayesian models for linked corpus. In *AAAI*, volume 10, page 1.

Saurabh Kataria, Prasenjit Mitra, Cornelia Caragea, and C. Lee Giles. 2011. Context sensitive topic models for author influence in document networks. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2274–2280.

Marijn Koolen and Jaap Kamps. 2010. The importance of anchor text for ad hoc search revisited. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 122–129, New York, NY, USA. ACM.

Reiner Kraft and Jason Zien. 2004. Mining anchor text for query refinement. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 666–674, New York, NY, USA. ACM.

Wendy Lehnert, Claire Cardie, and Ellen Rilofl. 1990. Analyzing research papers using citation sentences. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, pages 511–518.

David D. Lewis and Marc Ringuette. 1994. A comparison of two learning algorithms for text categorization. In *In Third Annual Symposium on Document Analysis and Information Retrieval*, pages 81–93.

Qing Lu and Lise Getoor. 2003. Link-based classification. In *International Conference on Machine Learning*, pages 496–503.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Qiaozhu Mei and ChengXiang Zhai. 2008. Generating impact-based summaries for scientific literature. In *Proceedings of ACL-08: HLT*, pages 816–824, Columbus, Ohio.

Donald Metzler, Jasmine Novak, Hang Cui, and Srihari Reddy. 2009. Building enriched document representations using aggregated anchor text. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 219–226, New York, NY, USA. ACM.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proc. of the 22nd Intl. Conference on Computational Linguistics*, COLING '08, pages 689–696, Manchester, United Kingdom.

Vahed Qazvinian, Dragomir R. Radev, and Arzucan Özgür. 2010. Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 895–903.

Anna Ritchie, Simone Teufel, and Stephen Robertson. 2006. How to find better index terms through citations. In *Proc. of the Workshop on How Can Computational Linguistics Improve Information Retrieval?*, CLIIR '06, pages 25–32, Sydney, Australia.

Anna Ritchie, Stephen Robertson, and Simone Teufel. 2008. Comparing citation contexts for information retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 213–222, New York, NY, USA. ACM.

Xiaolin Shi, Jure Leskovec, and Daniel A. McFarland. 2010. Citing for high impact. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 49–58.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 103–110.

S. Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, University of Edinburgh,.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 235–243, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1:67–88.

Wen Zhang, Taketoshi Yoshida, and Xijin Tang. 2011. A comparative study of tf* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765.

Xiaojin Zhu. 2005. Semi-Supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison.