Motif-Based Hyponym Relation Extraction from Wikipedia Hyperlinks

Bifan Wei, Jun Liu, Jian Ma, Qinghua Zheng, Wei Zhang, and Boqin Feng

Abstract—Discovering hyponym relations among domain-specific terms is a fundamental task in taxonomy learning and knowledge acquisition. However, the great diversity of various domain corpora and the lack of labeled training sets make this task very challenging for conventional methods that are based on text content. The hyperlink structure of Wikipedia article pages was found to contain recurring network motifs in this study, indicating the probability of a hyperlink being a hyponym hyperlink. Hence, a novel hyponym relation extraction approach based on the network motifs of Wikipedia hyperlinks was proposed. This approach automatically constructs motif-based features from the hyperlink structure of a domain; every hyperlink is mapped to a 13-dimensional feature vector based on the 13 types of three-node motifs. The approach extracts structural information from Wikipedia and heuristically creates a labeled training set. Classification models were determined from the training sets for hyponym relation extraction. Two experiments were conducted to validate our approach based on seven domain-specific datasets obtained from Wikipedia. The first experiment, which utilized manually labeled data, verified the effectiveness of the motif-based features. The second experiment, which utilized an automatically labeled training set of different domains, showed that the proposed approach performs better than the approach based on lexico-syntactic patterns and achieves comparable result to the approach based on textual features. Experimental results show the practicability and fairly good domain scalability of the proposed approach.

Index Terms—Hyponym relation extraction, wikipedia hyperlink, network motif

1 INTRODUCTION

The term hyponym indicates “a-type-of” relationship [9]. For example, maple is a hyponym of tree, and tree is a hyponym of plant. If lexical term ti is the hyponym of another lexical term tj, then ti and tj have a hyponym relation. The hyponym relation is a fundamental type of semantic relation connecting a wide variety of concepts or domain-specific terms to form a semantic taxonomy [33]. Hyponym relation extraction from Web pages plays a crucial role in web information extraction [7], taxonomy learning [25], knowledge acquisition [2], and other knowledge-rich problems.

Wikipedia has become a popular data source in hyponym relation extraction research. Several such studies adopted the syntactic-pattern-based methods or textual-feature-based machine learning methods [3], [5], [19], [20], [29], [30] to extract hyponym relations from Wikipedia. These methods rely mainly on features extracted from the text content of Wikipedia. When shifting to a new domain, these methods require new syntactic patterns to be learned or new training samples to be manually constructed, which usually entail high labor costs. In addition, these methods do not fully utilize the topological structure of hyperlinks in Wikipedia article pages.

1.1 Wikipedia Article Graph

This paper considers the topological structure of Wikipedia hyperlinks as an important type of feature in hyponym relation extraction. Each Wikipedia article page represents a domain-specific term. It contains a number of hyperlinks pointing to other article pages. For example, in Fig. 1 the first hyperlink indicates that a domain-specific term denotes that "maple" is a hyponym of "tree," and the second hyperlink indicates that "plant" is a hyponym of "tree." Hence, a novel hyponym relation extraction approach based on the network motifs of Wikipedia hyperlinks was proposed. This approach automatically constructs motif-based features from the hyperlink structure of a domain; every hyperlink is mapped to a 13-dimensional feature vector based on the 13 types of three-node motifs. The approach extracts structural information from Wikipedia and heuristically creates a labeled training set. Classification models were determined from the training sets for hyponym relation extraction. Two experiments were conducted to validate our approach based on seven domain-specific datasets obtained from Wikipedia. The first experiment, which utilized manually labeled data, verified the effectiveness of the motif-based features. The second experiment, which utilized an automatically labeled training set of different domains, showed that the proposed approach performs better than the approach based on lexico-syntactic patterns and achieves comparable result to the approach based on textual features. Experimental results show the practicability and fairly good domain scalability of the proposed approach.

1.2 Two Observations from WAGs

This paper explores the connectivity patterns in the WAGs to discover the hyponym relations based on the two observations below.

Our first observation is that the WAG of a domain contains many frequently recurring connectivity patterns known as network motifs in [18]. Network motifs (abbreviated to motifs) appear much more frequently in a specific
network than in randomized networks. They are suggested to be the fundamental building blocks carrying information-processing functions [1]. Furthermore, we find that hyponym hyperlinks are more likely to appear in some types of motifs than in other types [31]. Based on these observations, we extract motif-based features for classification model learning. The trained model is used to classify different types of hyperlinks. Therefore, the construction of features from local topological properties for hyponym relation extraction is the key problem in this study.

Second, we find that the features extracted from network motifs are closely related to the corresponding domain-specific article graph. The classification model acquired from the motif-based features for one domain cannot be directly applied to another domain in hyponym relation extraction. For this reason, generating training sets with minimal human involvement for extraction in the different domains is a challenging task.

1.3 Overview of Our Approach

To solve the two problems mentioned above, this paper proposes a novel extraction approach called MOtif-based hyponym Relation Extraction (MORE) for hyponym relation discovery. MORE focuses on hyponym relation discovery from academic domains in Wikipedia, such as Data mining, Classical mechanics and Microbiology. What MORE does includes:

1) automatically construct motif-based features from the WAG of a domain;
2) extract structural information from Wikipedia and heuristically label the training set of the domain based on the extracted structural information;
3) learn the classification model from the training sets to discover hyponym relations.

We conducted two experiments to validate the approach with seven domain-specific datasets from Wikipedia. One experiment used manually labeled data. The other used automatically labeled data. The experimental results show that the proposed three-node motif-based approach outperforms the approach based on lexico-syntactic patterns [9], and is on a par with the approach based on textual features [11], [34] in terms of F1 score. For example, the average F1 score of MORE with Random Forest in the second experiment outperforms the best result of the baseline experiment based on textual features by 9%.

1.4 Contributions and Organization

The main contributions of this study are as follows:

1) An automatic feature-constructing method based on network motifs is proposed for hyponym relation extraction. The effectiveness of the motif-based features is validated with seven real-world datasets of diverse domains. This type of feature can be directly utilized by classifiers and can achieve satisfactory performance without combining textural features.
2) A new training set labeling method based on weak supervision is proposed for hyponym relation extraction. We find that the navigation box, a type of structural information in Wikipedia article pages, contains plenty of domain-specific hyponym relations that can be automatically extracted. Navigation boxes and the Wikipedia category structure extracted from Wikipedia category pages (abbreviated to category pages or categories) are explored to label training sets with a heuristic algorithm.

The rest of the paper is organized as follows. Section 2 discusses related literature. Section 3 analyzes the three-node motifs of hyperlinks and provides a method to construct feature vectors from the three-node motifs. Section 4 describes the extraction of structural information from Wikipedia and illustrates the automatic labeling method to build training sets. Section 5 discusses the experimental results, and Section 6 presents the conclusion and suggestions for future research.

2 Related Work

Hyponym relation extraction from a text corpus has long been studied by taxonomy induction, statistical relational learning, and other research communities. Syntactic-pattern-based extraction and textual-feature-based classification methods are the two types of conventional methods mentioned in related literature.

Syntactic-pattern-based methods usually employ lexico-syntactic patterns, either handcrafted or automatically induced from training data, to identify hyponym relations. Hearst’s research [9] has played an important role in this subfield. Hearst reported six basic lexico-syntactic patterns for hyponym relation extraction and several techniques to obtain new patterns. Other researchers extended lexico-syntactic patterns in different aspects based on Hearst’s seminal work. For example, Snow et al. [28] proposed a method to automatically extract additional lexico-syntactic patterns from WordNet and a large text corpus by analyzing the dependency path of English sentences. Kozareva et al. [14] extracted hyponym words with hyponym pattern linkage graphs created from the lexico-syntactic pattern “CLASS NAME such as CLASS MEMBER and *.”

Syntactic-pattern-based methods are effective and reliable and can achieve relatively high precision [21] when large amounts of domain corpora are available, such as Wikipedia. However, these methods generally suffer from low recall. Furthermore, domain experts need to spend

1. The F1 score is the harmonic mean of precision and recall defined as the number of correctly identified hyponym relations divided by the extracted number and the golden standard number, respectively.
plenty of time analyzing sentences and summarizing patterns.

Textual-feature-based classification methods rely on the various features of text content to discover hyponym relations. Popular features include semantic similarity (or distance), term co-occurrence, link co-occurrence, text format, and syntactic and lexical-semantic information of sentences. Kambhatla [11] employed a maximum entropy classifier by combining Words, Overlay, Parse Tree and other textual features for relation extraction. Zhou et al. [34] improved the work of Kambhatla [11] by combining more textual features such as Base noun phrase features and Semantic resource, and employed a support vector machine (SVM) classifier for relation extraction. The subsumption algorithm [26] proposed by Sanderson and Croft assumes that a term pair (x, y) that meets the restrictions of $P(x|y) > 0.8$ and $P(y|x) < 1$ has a hyponym relation. Shinzato and Torisawa [27] employed lists in HTML documents to extract hyponym relations. The candidate hyponym relations were filtered by document frequencies and verb-noun co-occurrences.

Classification methods based on textual features have higher recall than syntactic-pattern-based methods. However, the former relies heavily on the carefully selected training set.

Massive user-generated tags have become available in recent years with the adoption of Web 2.0 technologies by many Web sites. The Wikipedia category system (WCS) is a high-quality tagging system because many people collaboratively maintain these tags. Several researchers have begun to leverage these tags in semantic relation discovery. Choiet et al. [6] formulated three hypotheses based on the lexical properties of the Wikipedia category structure and Wikipedia article pages. The researchers proposed a graph-based approach based on these hypotheses to extract taxonomic relations from Wikipedia category structures with the use of lexical topic words. Ponzetto and Strube [22] discovered the semantic relations between category titles based on the syntactic patterns of category titles, hierarchical patterns of WCS, and transitivity of relations. Kotlerman et al. [13] proposed a semi-automatic method based on WCS to the construct taxonomy of the videos in a target domain. They first constructed a domain-specific directed acyclic graph (DAG) from WCS. Then they pruned the DAG to retain only the most relevant categories to form the taxonomy of the target domain.

All these methods rely mainly on textual features extracted from text content. In contrast, the approach proposed in the present study relies only on local topological features extracted from the hyperlink structure. Moreover, a weakly supervised approach was proposed to heuristically label a training set by employing multiple structural information extracted from Wikipedia.

### 3 Feature Construction Based on Three-Node Motifs

The network motifs of network $N$ are connectivity patterns that occur far more often in $N$ than in randomized networks with the same degree distribution [12]. Several researchers studied the network motifs of an entire graph extracted from Wikipedia [10], [32], [35]. However, to the best of our knowledge, no attempt has been made to analyze the network motifs of a domain-specific WAG.

According to the number of nodes, network motifs are categorized into different types, such as three-node, four-node, and five-node motifs. Three-node motifs play a fundamental role in motif-based research. The structures of all possible three-node motifs can be enumerated easily by programs and observed visually by humans. The analysis of three-node motifs has always been the starting point of research [17], [24], [35]. The real-world datasets in the present study indicate that performance (F1 score of classification) based on three-node motifs is similar to that based on four-node or five-node motifs (more details can be found in Appendix 2, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2013.183). However, the overhead computation in three-node motif discovery is less than that in four-node or five-node motif discovery. For example, the running time to discover three-node motifs within the seven datasets is in the range of 28 s to 4 m, whereas the running time to discover five-node motifs within the same datasets is in the range of 2 h to 270 h. All calculations were performed on a high-performance server with two 2.40 GHz 10 core processors and 64 GB DDR3 RAM. Thus, this study focuses only on automatic feature construction based on three-node network motifs.

Three-node network motifs consist of 13 nonisomorphic connectivity patterns [18] as shown in Table 1. The $th$ motif in the table is referred to as motif $j$ ($1 ≤ j ≤ 13$) in the succeeding sections.

Motif 3 is regarded as an example in Fig. 2. A, B, and C represent the three domain-specific terms (article pages). The arrows are the directed hyperlinks in these pages. If $B$ is a hyponym of $A$, then the hyperlink between $B$ and $A$ is a hyponym hyperlink. Otherwise, if $B$ and $A$ have a connection but not a hyponym relation, then the hyperlink between $B$ and $A$ is an other hyperlink. In this example, $B$: Metadata is a hyponym of $A$: Data, whereas $A$: Data has no hyponym.

<table>
<thead>
<tr>
<th>ID</th>
<th>Connectivity Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>

### Table 1

<table>
<thead>
<tr>
<th>ID</th>
<th>Connectivity Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>
relation with C: Data warehouse. Thus, the two hyperlinks connecting A and B are labeled as hyponym relation, and the remaining two hyperlinks are labeled as other relation.

3.1 Motif Analysis of Wikipedia Hyperlinks

The Wikipedia hyperlinks of the seven domains related to Data mining, Computer network, Data structure, Euclidean geometry, Classical mechanics, Microbiology, and Wine were carefully scrutinized. The reasons why we chose the first three domains are: (1) these three domains are initially analyzed, and served as the source of our inspiration about MORE; (2) we have extensive knowledge about these three domains coming from the computer science discipline, which help facilitates the evaluation on those domains. The other four domains coming from distinct disciplines were utilized to verify the generalization and robustness of MORE.

The construction of the seven domain-specific datasets from the Wikipedia hyperlinks is described below.

1) For each of the seven domains, we crawl the Wikipedia article pages from the start position to a depth of 3. With the domain Data mining as an example, we crawl the article pages by traversing article-article hyperlinks from the Data mining article page.4

2) A set of URL regular expressions was utilized during crawling to remove irrelevant article pages. A great number of three-node motifs were identified in the 3,141 hyperlinks. The percentage of hyponym hyperlinks in this dataset is approximately 0.18 (=559/3,141). Table 2 shows that the HHRs of the motifs are different from one another (0.412, 0.361, and 0.407, respectively). The HHRs are all higher than 0.18, which shows the high level of hyponym hyperlinks in the motif are. This condition means that if a hyperlink appears in a motif with high HHR, then this hyperlink is likely to be a hyponym hyperlink.

Table 3 lists some statistics of the seven domain-specific datasets. The column "#Instances" indicates the total number of three-node motif instances.

The two parameters below were utilized to qualify the three-node motifs.


1) Z-Score indicates the statistical significance of a network motif [18]. The Z-Score of motif j is formally defined in (1).

\[
Z\text{-Score}(j) = \frac{N(j) - \bar{N}_R(j)}{\sigma_R(j)}.
\]  

where \(N(j)\) is the number of occurrences of motif \(j\) (1 \(\leq j \leq 13\)) in network \(N\). \(\bar{N}_R(j)\) is the average number of occurrences of motif \(j\) in an ensemble of randomized networks with the same degree of distribution as network \(N\). \(\sigma_R(j)\) is the standard deviation of \(N_R(j)\). In general, a motif with a high Z-Score indicates that the motif appears in a particular network (\(N\)) more frequently than in randomized networks.

2) A new parameter, Hyponym Hyperlink Rate (HHR), was introduced to describe the sparsity of hyponym relations within a network motif. The HHR of motif \(j\) is defined in (2). The higher the HHR of a network motif is, the denser the hyponym hyperlinks in the motif are. This condition means that if a hyperlink appears in a motif with high HHR, then this hyperlink is likely to be a hyponym hyperlink.

\[
HHR(j) = \frac{\# \text{ hyponym hyperlinks contained by the instances of motif } j}{\# \text{ all hyperlinks contained by the instances of motif } j}.
\]  

Table 3 illustrates several significant three-node motifs in the Data mining dataset.

The Instance column shows the examples of corresponding motifs. With motif 3 as an example, the instance is (A: Data, B: Metadata, C: Data warehouse), which is also explained with Fig. 2 in detail.

Two findings were discovered by analyzing the three-node motif instances of the different domains. The findings were utilized to characterize the three-node motifs.

After analyzing the hyponym hyperlinks in each motif and enumerating the instances of this motif, we find that the HHRs of the motifs are different from one another and from the percentage of hyponym hyperlinks in the entire network. With Data mining as an example, 559 hyponym hyperlinks were identified in the 3,141 hyperlinks. The percentage of hyponym hyperlinks in this dataset is approximately 0.18 (=559/3,141). Table 3 shows that the HHRs of motif 1, 2, and 3 are different from one another (0.412, 0.361, and 0.407, respectively). The HHRs are all higher than 0.18. This example means that if a hyperlink appears in motif

Fig. 2. Connectivity pattern and instance of motif 3.
1, 2 or 3, then this hyperlink is a hyponym hyperlink with different likelihoods.

Analysis of the other six datasets shows similar results. Thus, we conclude the first finding that different motifs indicate different probabilities of a hyperlink being a hyponym hyperlink.

The HHR and the Z-Score of a network motif were also found to be positively correlated. For ease of comparison, the Z-Scores were normalized by (3) as follows:

\[
\text{Normalized Z-Score}(j) = \frac{Z\text{-Score}(j)}{\sqrt{\sum_{k=1}^{13} Z\text{-Score}^2(k)}}. \tag{3}
\]

The normalized Z-Scores and the HHRs of the three-node motifs in the Data mining dataset are plotted in Fig. 3. Uncertainty exists as to whether correlation computing is reliable when working with small dimensions (i.e., less than 15). Bootstrapping method was therefore utilized to produce 1,000 samples, and the correlation between Z-Score and HHR was computed based on the samples. The minimum value is positive for the seven datasets. Most of the correlation coefficients are in the range of 0.7 to 1, indicating that the relationship between Z-Score and HHR is not accidental.

Analysis of the other six datasets shows similar results (more details can be found in Appendix 1, available in the online supplemental material). Thus, we conclude the second finding that the HHRs and Z-Scores of the three-node motifs are positively correlated.

Network motifs were leveraged based on the two findings to construct discriminative features in the discovery of hyponym relations from the structure of the hyperlinks.

### 3.2 Motif-Based Feature Construction

MORE is proposed based on the above observations for automatic motif-based feature construction from the WAG of a domain. MORE learns the classification models from the training set by utilizing motif-based features.

A WAG is formally represented as a directed graph, \( G = (V, E) \), where \( V \) is a set of vertices. Each vertex represents an article page. \( E \) is the set of directed edges between two article pages (one per directed hyperlink). An edge and its adjacent edges form a set of network motifs that can be utilized by MORE as features to identify the semantic type of this particular edge. Feature construction can be further divided into three steps [31].

First, MORE discovers all instances of three-node motifs from a domain-specific WAG.

Second, MORE automatically enumerates all hyperlinks in every motif instance. Each hyperlink is regarded as an analysis target. This procedure is necessary because a hyperlink may simultaneously appear in multiple instances of a network motif. MORE calculates the occurrences in different network motifs for each hyperlink. \( O(i,j) \) represents the number of times hyperlink \( i \) appears in the instances of motif \( j \) as shown in Table 1. The occurrences indicate the impact of the strength of this type of network motif.

A part of a WAG is considered an example in our experiment as shown in Fig. 4. With hyperlink \( E1 \) as the target, hyperlink \( E1, E5, \) and \( E6 \) form one instance of motif 1. Simultaneously, hyperlink \( E1, E4, E7, \) and \( E8 \) form three instances of motif 5. Hence, \( O(1,1) \) and \( O(1,5) = 3 \).

Third, MORE builds a feature vector for every hyperlink within the WAG based on the occurrences of different network motifs. The HHR of a motif reveals the probability of a hyperlink being a hyponym hyperlink. The occurrences of the network motifs can be weighted by the HHRs of the corresponding network motifs. Hence, the weighted vector component \( F(i,j) \) for hyperlink \( i \) and motif \( j \) can be expressed as follows:

\[
F(i,j) = O(i,j) \times \text{HHR}(i). \tag{4}
\]

However, the HHR of a motif cannot be achieved prior to labeling all the hyperlinks. Fortunately, the Z-Score of a three-node motif is positively correlated with the HHR of the same motif and can be calculated from the network structure without labeling all the hyperlinks. For this reason, Z-Score was utilized as the weight of a motif instead of HHR. Equation (4) can be reformulated as follows:

\[
F(i,j) = O(i,j) \times \text{Z-Score}(i). \tag{5}
\]

\( \text{Z-Score}(i) \) is defined in (1). Empirical experiments show that the performance using (5) is similar to that using (4) (more details can be found in Appendix 4, available in the online supplemental material).

The feature vector of hyperlink \( i \) is presented as follows:

\[
FV(i) = \{F(i,1), F(i,2), \cdots , F(i,13)\}. \tag{6}
\]

All the thirteen types of three-node motifs with their Z-scores as weights are exploited to construct discriminative features.

MORE transforms the hyponym relation extraction problem into a binary classification problem with these
motif-based features. Every hyperlink connecting a pair of domain-specific terms is represented by feature vector \( x_i \in X \subseteq \mathbb{R}^{13} \). The hyponym relation extraction problem is expressed as binary classification model \( f: X \rightarrow Y \) acquired from training set \( TS = \{ t_i \} \), where \( t_i = (x_i, y_i) \) and \( y_i \in Y = \{ \text{hyponym, other} \} \). The classification models acquired from the training set can perform fairly well in the corresponding test set within one domain.

However, a classification model acquired from one domain cannot be directly applied to other domains in hyponym relation extraction. When a classification model trained in domain \( A \) is utilized in domain \( B \), its performance declines significantly. Training set generation becomes a problem when a new domain emerges.

4 TRAINING SET LABELING BASED ON WEAK SUPERVISION

A weakly supervised learning method was leveraged to acquire a training set for every domain. The following two challenges were addressed in the automatic labeling of training data.

1) Discovery and extraction of high-quality structural information at minimum costs;

2) Leverage of a heuristic algorithm with the extracted information to label training sets.

In addressing the first challenge, we found that the Wikipedia navigation box, a type of structural information in Wikipedia article pages, contains plenty of domain-specific hyponym relations that can be automatically extracted. This type of structural information is suitable for automatic labeling of training data with weakly supervised learning because of its high quality.

A heuristic algorithm, Automatic Training Set Labeling (ATSL), was developed for the second challenge to label training sets with navigation boxes and WCS. The result of the proposed method meets the desired objectives compared with manual annotation method.

In summary, MORE automatically extracts domain-specific hyponym relations from navigation boxes within Wikipedia article pages and category pages in WCS to minimize human labor cost. The training set of every domain is then heuristically labeled.

Several important definitions of the two types of structural information, including necessary pre-processing of WCS, are introduced in this section. The heuristic labeling algorithm is also described with the two types of structural information.

4.1 Navigation Boxes

A navigation box (Navbox) is a table at the bottom or the right-hand side of a Wikipedia article page. A Navbox helps users navigate through a set of closely related Wikipedia article pages.

More than 2 million English article pages contain at least one Navbox,6 accounting for almost 50% of the 4.2 million English Wikipedia article pages.7 A Navbox represents the hyponym relations among article pages within the corresponding table. The caption of the table represents the root article page of the table. The header cell in every row is the subordinate article page of the root article page and is also a superordinate article page of other article pages in the same row. The Navbox in the Regression analysis page is shown in Fig. 5. Linear regression is the heading of a line. The elements of this line, including Simple linear regression, Ordinary least squares and Generalized least squares, are all hyponyms of Linear regression.

A Navbox tree is a tree built automatically by MORE from a Navbox of a Wikipedia article page. The nodes of this tree represent the article pages in the Navbox. The root of the tree is the table caption of the Navbox. A Navbox tree represents the hyponym relations of all article pages in a box. If an article page is a child node of another article page in a Navbox tree, then these two article pages are likely to have a hyponym relation because Navboxes are created by Wikipedia editors to help users navigate through article pages with hyponym relations. We confirmed this assumption by manually checking the Navbox links of the seven datasets. For example, 92.7% of the Navbox links we verified in the Data mining domain represent hyponym relations.

The Navbox Forest of a hyperlink is defined as a collection of Navbox trees in two article pages connected by this hyperlink. For example, the Navbox Forest of the hyperlink from the Data mining article page to the Regression analysis article page includes two Navbox trees: Navbox Data warehouse and Navbox Regression analysis.

Navboxes provide relatively reliable domain-specific hyponym relations and can be utilized to label the positive instances of a training set. However, only a small number of hyponym relations are covered by Navboxes.

4.2 Wikipedia Category System

Wikipedia Category System (WCS) is another good source for building high-quality heuristics-based training set. WCS embeds smaller taxonomies suitable for specific domains [13]. A set of tree structures was constructed for automatic data labeling from WCS by performing two pre-processing tasks sequentially.

The first pre-processing task is to select a collection of domains as the related domains of the target domain. For example, the “related domains” of Data mining include Database system, Statistics, Machine learning, Information science, and Visualization. Every domain in Wikipedia corresponds to a category called root category. The title of a root category is usually the name of this domain. For example,
Data mining\textsuperscript{8} is the root category of the domain Data mining. The “related domains” of a domain are acquired by a semi-automatic method that does the following:

1) analyze the outgoing hyperlinks of a root article page and obtain all the target article pages;
2) parse all the target article pages and determine the categories in these article pages;
3) domain experts manually select several categories as the root categories of “related domains.”

The second pre-processing task is to build a category tree for every domain from WCS, which has a non-strict tree structure. The building process is divided into the following two steps based on the results of the first task.

1) From the root category of a domain, MORE traverses the category-category hyperlinks in a breadth-first order. Fig. 6 illustrates the root category of the Data mining domain, which contains three sections: Subcategories, Pages in category and Categories section. The traversed category-category hyperlinks are located in the Subcategories section.

2) MORE combines the traversed categories and category–category hyperlinks to form a category tree. The internal nodes and leaf nodes of the category tree represent the category pages in the Subcategories section and the Pages in category section, respectively.

From a domain and its closely related domains, MORE builds a collection of category trees called a Category Forest. For example, the category forest of the domain Data mining includes the tree itself and the category trees of the related domains, namely, Database system, Statistics, Machine learning, Information science, and Visualization.

Fig. 7 presents the WAG and category forest of a domain. The WAG of the domain is shown in the middle of the figure. The corresponding category forest, which is on the left and right side of the WAG, consists of two category trees. The tree on the left is the category tree of the domain, and the one on the right is the category tree of a related domain.

Fig. 7 presents the four different types of edges, described as follows with the aid of Fig. 6.

- A solid line with double arrows indicates a category-category hyperlink, which is a type of bidirectional hyperlink connecting two category pages. For example, the hyperlink from the Subcategories section (shown in Fig. 6) of category page Data mining to category page Cluster analysis and the hyperlink from the Categories section of article page Anomaly detection to Data mining form a category-category hyperlink.
- A dotted line with double arrows indicates a category-article hyperlink. The hyperlink from the pages in category section (shown in Fig. 6) of category page Data mining to article page Anomaly detection and the hyperlink from the Categories section of article page Anomaly detection to category page Data mining offer a typical example of a category-article hyperlink.
- A dotted line tagged as “E” indicates the Main Article hyperlink, which is a special category–article hyperlink shown in Fig. 7. The category and corresponding article pages have the same title. For example, the Main Article of the category Data mining is the article page Data mining.
- A solid line with only one arrow denotes a hyperlink between two Wikipedia article pages.

Fig. 7 shows that C3 is a subordinate category of C1; A1 and A2 are the main articles of category C1 and C3, respectively; and A3 is a product of category C3. The following can be derived from this category structure.

1) Article page A2 is a subordinate article of A1; this means that A2 is a hyponym of A1.
2) Article page A3 is a subordinate article of A2; this means that A3 is a hyponym of A2.

Based on this example, we can further generalize that if x is a subordinate article of article page y, the hyperlink between x and y is labeled as hyponym relation. Otherwise, the hyperlink should be labeled as other relation.

The category pages in WCS contain many domain-specific hyponym relations that are complementary to the hyponym relations contained in the Navboxes.

4.3 Automatic Labeling

MORE heuristically labels the training set of a dataset based on the automatically extracted structural information, including the Navbox forest of a hyperlink and the category forest from WCS. The automatic labeling process is described with ATSL as follows:

\textsuperscript{8} http://en.wikipedia.org/wiki/Category:Data\_mining
Algorithm ATSL() //automatic training set labeling
Input: ES; //the hyperlink set of a domain, represented
        //by the edge set in the WAG of the domain
Output: TSp, TSn; //positive and negative training set
1: Extracting Category Forest and Navbox Forest;
2: TSp = ∅; TSn = ∅;
3: Selecting an edge e = ⟨vx, vy⟩ ∈ ES;
   ES = ES − {e};
4: If RDCF(vx, vy) = ES
   TSp = TSp ∪ {e}; //positive set;
5: If RDCF(vx, vy) = hyponym relation
   TSp = TSp ∪ {e}; //positive set;
   Else if RDCF(vx, vy) = other relation
   TSn = TSn ∪ {e}; //negative set;
6: If |TSp ∪ TSn| < σ Goto STEP3; //typically, σ = 30% × |ES|
7: Return TSp, TSn.

The algorithm Relation Detection by Navbox Forest (RDNF) in ATSL detects whether two domain-specific terms in the same Navbox forest have a hyponym relation. The coverage of Navbox forests is small. If two domain-specific terms do not exist in the same Navbox forest, determining whether the two terms have a hyponym relation is difficult. Thus, only the Navbox forest of a hyperlink is utilized to label positive instances.

Algorithm RDNF() //relation detection by navbox forest
Input: e = ⟨vx, vy⟩; //an edge
        NavboxForest; //navbox forest
Output: True or False;
1: For (each tree t in NavboxForest)
   If Descendant(vx, vy) or Descendant(vy, vx)
      Return True;
2: Return False.

In RDNF, Descendant(vx, vy) determines whether vertex vx is a descendant of vertex vy in a tree t.

The purpose of the algorithm Relation Detection by Category Forest (RDCF) in ATSL is to label an edge by means of a category forest. RDCF heuristically labels an edge e = ⟨vx, vy⟩ as one of three types: hyponym, other relation or unknown. Unknown indicates that the taxonomic relation of a hyperlink cannot be identified by the category forest.

The three fundamental functions utilized in RDCF are explained below.
1) Descendant(vx, vy) is similar to the function with the same name in RDNF.
2) LowestCommonAncestor(vx, vy) computes the lowest common ancestor of vx and vy.
3) PathLength(vx, vy) computes the path length between vx and vy.

The automatic and manual labeling results were compared as shown in Table 4.

It is seen from Table 4 that the average precision of automatic labeling is 85.6% and average recall is 42.9%, which means that automatic labeling method can label a considerable amount of data with high precision. In practice, satisfactory results can be achieved when 30% of the hyperlinks of the dataset are randomly selected as the training set.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Hyperlinks</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data mining</td>
<td>3,141</td>
<td>0.893</td>
<td>0.323</td>
</tr>
<tr>
<td>Computer network</td>
<td>3,079</td>
<td>0.826</td>
<td>0.395</td>
</tr>
<tr>
<td>Data structure</td>
<td>3,081</td>
<td>0.884</td>
<td>0.498</td>
</tr>
<tr>
<td>Euclidean geometry</td>
<td>10,369</td>
<td>0.898</td>
<td>0.501</td>
</tr>
<tr>
<td>Classical mechanics</td>
<td>5,185</td>
<td>0.864</td>
<td>0.459</td>
</tr>
<tr>
<td>Microbiology</td>
<td>9,406</td>
<td>0.801</td>
<td>0.352</td>
</tr>
<tr>
<td>Wine</td>
<td>2,116</td>
<td>0.826</td>
<td>0.472</td>
</tr>
</tbody>
</table>

Algorithm RDCF() //relation detection by category forest
Input: e = ⟨vx, vy⟩; //an edge
        CategoryForest; //category forest
Output: Relation type;
1: MinDistance = ∞;
2: For (each tree t in CategoryForest)
   If Descendant(vx, vy) or Descendant(vy, vx)
      Return hyponym relation;
   Else lca = LowestCommonAncestor(vx, vy);
      d = PathLength(lca, vx) + PathLength(lca, vy);
      If MinDistance > d MinDistance = d;
3: If MinDistance > 2 Return other relation;
4: Return unknown.

The ATSL algorithm was utilized to determine if a hyperlink is an element of the training set. The detailed process is described as follows:

1) MORE randomly selects a hyperlink (connecting a pair of domain-specific terms) and leverages the ATSL algorithm to assign a label to this hyperlink. If the result is hyponym or other, this hyperlink will be considered an element of the training set.

2) MORE repeatedly performs process 1 until a training set that contains 30% of the hyperlinks of the dataset is achieved.

5 Experiments
With the motif-based features of hyperlinks and the automatic labeled training set, MORE transforms the hyponym relation extraction problem into a binary classification problem. Five typical machine learning algorithms from “basic” Naïve Bayes to “advanced” SVM were utilized to evaluate the motif-based features. Seven real-world datasets of various sizes were created to validate MORE.

The two baseline experiments, based on lexico-syntactic patterns and textural features respectively, were described. MORE was then evaluated in the two experiments. The first experiment evaluates the classification effectiveness of MORE with manually labeled data. The second experiment evaluates the generalization capability of MORE.

5.1 Baseline Experiment Based on Lexico-Syntactic Patterns
Lexico-syntactic pattern-based extraction methods are typical hyponym relation extraction methods that employ text tokens and syntactic structures to discover hyponym
Data mining, k-medians clustering is a cluster analysis algorithm. In statistics and data relation. For example, in the sentence “If the words around a hyperlink are matched by these patterns, the hyponym relation between two entities within a sentence. The lexical, syntactic, and semantic structures of a sentence provide numerous features to train classification models. The entities represented by the titles of Wikipedia article pages are often domain-specific and seldom appear in a sentence from another text resource. Designing a textual-feature-based hyponym relation extraction experiment to identify the relation between two entities in Wikipedia is a challenging task.

The second baseline experiment was conducted based on textual features following the procedures employed in two highly-cited studies [11], [34]. In the beginning, all English Wikipedia article pages were processed and more than 60 million sentences were obtained. Domain-specific terms were found to appear often in domain-related article pages and seldom in other domains. For this reason, only domain-related article pages obtained by traversing through article–article hyperlinks from the starting position of a domain were considered.

Dataset construction was divided into the following four steps with Data mining domain as an example.

1) Plain text was extracted from the article pages and cleaned by removing the navigation boxes, information boxes, and other similar elements. Duplicated text, very short text and the text containing only math symbols were also erased. A total of 7,835 text files were obtained from the Data mining domain.

2) The sentences from the plain text were acquired with the sentence splitter provided by Stanford CoreNLP library. The sentences were then cleaned by removing duplicates, very short, and incomplete sentences. A total of 604,708 sentences were obtained.

3) The term mentions in every sentence were recognized by combining the string regular matching and hyperlink information of the sentence. A total of 9,428 sentences containing at least two different terms were chose.

4) Any two term mentions were combined, and relation mentions were formed for every sentence. A total of 17,186 instances were created and labeled from the selected sentences.

Words, Overlay, and Base Noun Phrase features were selected based on the description provided by two studies [11], [34]. These selected features can achieve better performance than other features such as Semantic Resource features, which can only improve the F1 score by 1% [34]. Features such as Entity Type (e.g., Organization and Person) that do not exist in the datasets were removed.

Zhou et al. [34] and Nanda Kambhatla [11] selected SVM and maximum entropy model as classifiers for relation extraction, respectively. In the present paper, LibSVM11 and Maximum Entropy Modeling Toolkit12 were leveraged to extract hyponym relations from the seven datasets. The results are shown in Table 7. The two classification models are set with the following default parameters:

- LibSVM: $C = 1.0$, $\epsilon = 10^{-3}$, $\text{kernel} = \text{linear}$, $\gamma = 1/\sigma^2$;
- Maximum Entropy Modeling Toolkit: iterations $\sim 30$, parameter estimate method = L-BFGS.

![TABLE 5](http://en.wikipedia.org/wiki/K-medians_clustering)

<table>
<thead>
<tr>
<th>ID</th>
<th>Lexico-syntactic Patterns</th>
<th>ID</th>
<th>Lexico-syntactic Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NP2 such as NP1+</td>
<td>6</td>
<td>NP2 consist(s) of NP1</td>
</tr>
<tr>
<td>2</td>
<td>such NP2 as NP1+</td>
<td>7</td>
<td>NP1 (and</td>
</tr>
<tr>
<td>3</td>
<td>NP1 is (a</td>
<td>an) NP2</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>NP2 intud(e</td>
<td>es</td>
<td>ing) NP1</td>
</tr>
<tr>
<td>5</td>
<td>NP1, one of NP2</td>
<td>10</td>
<td>NP1 (in</td>
</tr>
</tbody>
</table>

NP1 represents a subordinate Noun Phrase (NP). NP2 represents a superordinate Noun Phrase.

![TABLE 6](http://en.wikipedia.org/wiki/K-medians_clustering)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data mining</td>
<td>0.698</td>
<td>0.259</td>
<td>0.378</td>
</tr>
<tr>
<td>Computer network</td>
<td>0.657</td>
<td>0.271</td>
<td>0.384</td>
</tr>
<tr>
<td>Data structure</td>
<td>0.683</td>
<td>0.264</td>
<td>0.381</td>
</tr>
<tr>
<td>Euclidean geometry</td>
<td>0.725</td>
<td>0.292</td>
<td>0.416</td>
</tr>
<tr>
<td>Classical mechanics</td>
<td>0.664</td>
<td>0.283</td>
<td>0.397</td>
</tr>
<tr>
<td>Microbiology</td>
<td>0.701</td>
<td>0.264</td>
<td>0.384</td>
</tr>
<tr>
<td>Wine</td>
<td>0.731</td>
<td>0.259</td>
<td>0.382</td>
</tr>
</tbody>
</table>

relations from natural language text. These methods were utilized as the baseline because they are very popular and effective in processing web pages [8], [9], [28]. The extraction process, which follows that of [9], is described below.

First, the context words of the hyperlinks in Wikipedia article pages are analyzed in detail. The two sentences before a hyperlink and one sentence after the hyperlink are automatically extracted to determine the lexico-syntactic patterns.

Second, after manual examination and statistical analysis of the extracted sentences, the 10 lexico-syntactic patterns with the highest coverage are selected for pattern matching as shown in Table 5. These lexico-syntactic patterns can be utilized to identify hyponym relations.

Third, pattern matching based on part-of-speech tags was conducted to identify the semantic type of hyperlinks. If the words around a hyperlink are matched by these patterns, then the hyperlink would likely have a hyponym relation. For example, in the sentence “In statistics and data mining, k-medians clustering is a cluster analysis algorithm,” the pattern is a/an indicates that k-medians clustering and cluster analysis have a hyponym relation with high probability. However, many hyponym relations are not recognized by these patterns because no explicit pattern words can be found around the hyperlinks.

Table 6 shows that the baseline method achieves relatively high precision in the seven datasets by utilizing the recognized lexico-syntactic patterns. The F1 scores are similar to those in [28].

5.2 Baseline Experiment Based on Textual Features

Conventional textual-feature-based hyponym relation extraction identifies the hyponym relation between two


Table 7 shows that the F1 scores obtained in the present study are slightly lower than those obtained in [34]. The main reason is that Automatic Content Extraction (ACE 2004) datasets contain plenty of additional annotation information such as Mention Type and Mention Level, which do not exist in our datasets.

5.3 Evaluation Based on Manually Annotated Data
Each hyperlink in the seven domain datasets was annotated with a semantic tag (hyponym or other) for experiential verification. A strict process was established to ensure the quality of the annotation.

1) We provided the annotators annotated examples and annotation guidelines.
2) Every two annotators were asked to label the same datasets independently.
3) The pairwise agreement established by two annotators was measured with Cohen’s kappa coefficient [4]. The agreement coefficient of the seven datasets must be at least 0.85, which is described as an “almost perfect agreement” by Landis and Koch [15]. When an agreement was not reached, a third annotator was asked to break the tie.

MORE adopts five popular classification algorithms, namely, Naïve Bayes, KNN, C4.5, SVM, and Random Forest, to extract hyponym relations. The following were utilized as standard default in the five algorithms:

- SVM: $C = 1.0$, $\epsilon = 10^{-3}$, $Kernel = RBF$, $\gamma = 1/\sigma^2$;
- KNN: $k = 3$;
- C4.5: confidence factor = 0.25, binary split = false;
- Random Forest: number of trees = 10.

Table 8 shows the F1 scores based on standard 10-fold cross-validation. The training sets were randomly sampled in the runs. The best F1 scores are written in bold. The column MORE+$x$ indicates that MORE utilizes Algorithm $x$ for the motif-based features. The F1 scores of the baseline experiments and MORE with the five classification algorithms are shown in Fig. 8. “MORE+NB” and “MORE+RF” in Fig. 8 are the abbreviations of “MORE+Naïve Bayes” and “MORE+Random Forest,” respectively. Baseline 1 in Fig. 8 indicates performance based on the lexico-syntactic patterns. Baseline 2 in Fig. 8 indicates the performance of the maximum entropy classifier based on the textual features, which is better than that of the SVM classifier in Table 7.

The following observations are drawn from the statistical data presented in Table 8 and Fig. 8.

1) The F1 scores of the classification algorithms, except for Naïve Bayes, are higher than those of the lexico-syntactic-based and textual feature-based methods. For example, the average F1 score of MORE with Random Forest is higher than that of maximum entropy classifier by 12%.

2) The F1 scores vary in the different classification algorithms. The Random Forest algorithm performs better than the other algorithms in terms of F1 scores. SVM exhibits very high precision. The performance of Naïve Bayes is the worst probably because of the unsatisfied conditional independence assumption of the features.
The observations from the experiment reveal that features based on network motifs are effective for hyponym relation extraction.

### 5.4 Evaluation Based on Automatically Labeled Data

The configuration of this experiment is similar to that of the third experiment. However, the training sets in this experiment were automatically labeled using the weak supervision method described in Section 4.3.

In the seven datasets, 30% of all the hyperlinks were utilized as the training set labeled with structural information. The other 70% was utilized as the testing set labeled manually. Fig. 9 shows the F1 scores of the two baseline experiments and MORE with the five classification algorithms.

The following observations are drawn from the statistical data presented in Table 9 and Fig. 9.

1) All the classification algorithms, except Naïve Bayes, have higher F1 scores than the scores in baseline 1. These F1 scores are at the same level as the scores in baseline 2. For example, the average F1 score of MORE with Random Forest is higher than the score obtained from the best baseline system (the maximum entropy classifier) using textual features by 9%.

2) The F1 scores of the classification algorithms are slightly lower than those in Section 5.3 because the automatically labeled training sets contain falsely labeled instances that mislead classification model training. In addition, the sizes of the training sets are also smaller than those in Section 5.3.

Given the results of the experiment, we conclude that the structural information extracted from Wikipedia is effective for heuristically labeling training sets.

The combination of the experimental results in Sections 5.3 and 5.4 shows that (1) motif-based features are effective for hyponym relation extraction, and (2) weakly supervised learning method is reliable in providing automatically labeled training sets with good quality. These results indicate that the performance of MORE is robust and consistent in various domains.

Compared with textual-feature-based methods, our approach can discover the hyponym relation between two domain-specific terms when they appear in different sentences, different paragraphs, or even in different article pages. This type of relations can hardly be discovered by the textual-feature-based methods, which require the two domain-specific terms appear in the same sentence [9], [11], [34] to build the features. For example, in the baseline experiments the hyponym relation between Decision tree learning and Statistical classification cannot be correctly discovered by the methods proposed in [9], [11], [34] to build the features. For example, in the baseline experiments the hyponym relation between Decision tree learning and Statistical classification cannot be correctly discovered by the methods proposed in [9], [11], [34] to build the features. Instead, our approach, on the other hand, discovered the hyponym relation between them based on the motif-based features constructed from the hyperlink structure among Decision tree learning, Statistical classification, ID3 algorithm and other related article pages.

### 6 Conclusion

Numerous studies have been performed to extract hyponym relations from text corpora. However, minimal research has been conducted to extract hyponym relations from Wikipedia article pages based on local topological properties. A novel approach called MORE is proposed in the present study to address this problem. The key contributions of our study can be summarized as follows:

1) We find that the local connectivity patterns of Wikipedia hyperlinks are effective features in extracting hyponym relations. Three-node motifs are utilized in MORE to construct the discriminative features of the classification algorithms. The effectiveness of the three-node motif-based features is evaluated based on seven real-world Wikipedia datasets.

2) We find the Wikipedia navigation boxes as well as the WCS contain numerous domain-specific
hyponym relations. The structural information from them is automatically extracted by MORE to heuristically label the training sets of various domains. The robustness of the heuristically labeling algorithm based on the navigation boxes and WCS are examined with seven-real-world Wikipedia datasets.

This paper focuses on extracting hyponym relations from academic domains in Wikipedia. Such Wikipedia data contain a certain amount of hyponym relations between domain-specific terms. We showed that the motif-based features are effective in hyponym relation extraction. The experiment results showed that our approach works well in domains where there are enough hyponym relations. This approach may not work well in a domain where the hyponym relations among domain-specific terms are very sparse, such as human individuals or companies.

We also propose a potential direction for future research. The features based on text content can be combined with network motifs to improve extraction performance. Comprehensive experiments can be conducted to combine these two kinds of features through co-training method based on probability topic models.

ACKNOWLEDGMENTS

The research was supported in part by the National Science Foundation of China under Grants 61173112, 61221063, and 61202184; National High Technology Research and Development Program of China under Grant 2012AA011003; National Key Technology R&D Program under Grant 2011BAK08B02. The authors are grateful to the anonymous reviewers’ comments. The current work is the extension and continuation of our previous work that has been published in a conference paper of ICONIP 2012 [31].

REFERENCES

Bifan Wei received the BS degree in Aerocraft Dynamics Engineering from Beijing University of Aeronautics and Astronautics, China, in 2000. He is currently working toward the PhD degree in computer science at Xi’an Jiaotong University, China. His research interests include web data mining, faceted search and taxonomy learning.

Jun Liu received the BS, MS, and PhD degrees from Xi’an Jiaotong University, China, in 1995, 1998, and 2004, respectively, all in computer science. He is currently a professor in the Department of Computer Science, Xi’an Jiaotong University. His main research interests include text mining, data mining and e-learning. He has published more than 40 research papers in various journals and conference proceedings.

Jian Ma received the BS degree in Information and Computing Science from Anhui University of Science and Technology, China, in 2011. He is currently working toward the MS degree in computer science at Xi’an Jiaotong University, China. His research interests include data mining and information extraction.

Qinghua Zheng received the BS degree in computer software in 1990, the MS degree in computer organization and architecture in 1993, and the PhD degree in system engineering in 1997 from Xi’an Jiaotong University, China. He is currently a professor in Xi’an Jiaotong University, and the dean of the Department of Computer Science. His research areas include multimedia distance education, and computer network security.

Wei Zhang is a software development engineer at the Search Experience team of Amazon.com, LLC., USA. In the past, he also worked at the Data Mining team of Microsoft Bing.com. He received a PhD degree in Computer Science from the University of Illinois at Chicago, USA. He also holds a master’s degree and a bachelor’s degree in computer science. His search interests include Web search engine front end system design, Web user interface, and text mining. He has published 14 research publications in various journals and conferences.

Boqin Feng is a professor in the Department of Computer Science, Xi’an Jiaotong University, China. He is the National Award for Distinguished Teacher (The Ministry of Education, 2003), the dean of Computer Teaching and Experiment Center in Xi’an Jiaotong University. He is a specialist in teaching methods of the fundamental computer technology. He has also authored and co-authored more than 80 research papers.