Web Unit Based Mining of Homepage Relationships

Aixin Sun*
School of Computer Science and Engineering
University of New South Wales, Sydney NSW 2052, Australia

Ee-Peng Lim
School of Computer Engineering
Nanyang Technological University, Singapore 639798

Abstract

Homepages usually describe important semantic information about conceptual or physical entities, and are hence the main targets for searching and browsing. To facilitate semantic based information retrieval (IR) at a Web site, homepages can be identified and classified under some pre-defined concepts and these concepts are then used in query or browsing criteria, e.g., finding professor homepages containing "information retrieval". In some Web sites, relationships may also exist among homepages. These relationship instances (also known as *homepage relationships*) enrich our knowledge about these Web sites and allow more expressive semantic based IR. In this paper, we investigate the features to be used in mining homepage relationships. We systematically develop different classes of inter-homepage features, namely, *navigation*, *relative-location*, and *common-item* features. We also propose deriving for each homepage a set of support pages so as to obtain richer and more complete content about the entity described by the homepage. The homepage together with its support pages are known to be a *Web unit*. By extracting inter-homepage features from Web units, our experiments on the WebKB dataset show that better homepage relationship mining accuracies can be achieved.

Keywords: Homepage relationship mining, Web Unit, Inter-homepage features.

1 Introduction

1.1 Motivation

World Wide Web (or Web) today is populated with large number of Web sites and Web pages. The OCLC's Web characterization initiative estimated that there were 3 million public Web sites and 1.4 billion Web pages in 2002 [17]. Each public Web site had an average of 441 Web pages. On the other hands, there are millions of users searching and browsing Web sites each day. The above numbers certainly complicates the tasks of information seeking on the Web. In an online survey, Broder proposed a taxonomy to classify a Web query to be *navigational*, *informational*, or *transactional* [1] . It was found that about 25%, 40% and 35% of Web queries are navigational, informational and transactional respectively. Among the Web pages, there are only a small subset of them that are often the targets of navigational and informational queries (estimated 65% of Web queries) and these are known as **homepages**. Homepages are Web pages that describe information about some entities or concepts such as people, courses, departments and books[1]. Due to their importance, homepage finding task has been included as one of the annual TREC meetings and several methods have been proposed [10](see Section 6).

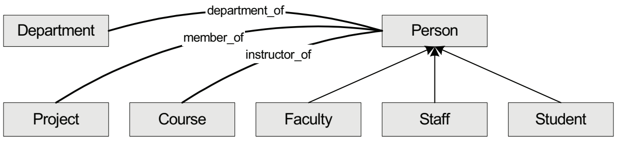

Figure 1: Concepts and relationships in university domain

When a Web site contains multiple homepages, these homepages may be assigned concept labels and they may be related by some relationships. Both these concepts and relationships can be a useful piece of information for querying the Web site. For example, in a university Web site, there are usually homepages of departments, professors, researchers, students, courses, and projects. As these homepages are related semantically by some relationships such as instructor-of(*professor, course*), member-of(*researcher, project*), and work-for (*professor, department*), it is possible to assign relationship labels to each pair of related homepages (see Figure 1 for the set of concepts and relationships appropriate to the university domain). With these homepages and homepage pairs properly labeled, Web queries can be augmented with concept and relationships to express more complex informational and navigational information needs. One can query, for example, the professors who teach some course about "data mining" or the researchers of the "Myriad" project.

In this paper, we call finding homepages from a Web site and assigning them concept labels the **homepage mining** problem,

*Aixin Sun is the corresponding author. The work was done when Aixin Sun was with the School of Computer Engineering, Nanyang Technological University, Singapore

[1] Note that the term homepage in this paper does not necessary refer to the homepage of a Web site. In other word, multiple homepages can be hosted by the same Web site.

and finding homepage pairs and assigning them relationship labels the **homepage relationship mining** problem. The former is different from the TREC's *homepage finding* problem in the following aspects.

- Homepage finding problem returns homepages relevant to some given user queries while homepage mining problem does not involve any user queries.
- The homepages returned by homepage finding problem are from a large collection of Web pages from different Web sites. The homepage mining problem, in contrast, focuses on pages from a single Web site.
- Concept labels are assigned to homepages returned by homepage mining but not homepage finding.

In our earlier work, key Web pages (or homepages) are identified and assigned concept labels using a Web unit mining method [18]. There are also other approaches to address homepage mining using Web classification techniques [12, 19].

The homepage relationship mining problem, on the other hand, is a relatively new research problem. To the best of our knowledge, only the CMU Text Learning group has proposed a rule-based classification method for assigning relationship labels to Web page pairs, as opposed to homepage pairs [5]. In other words, the notion of homepage was not considered by their research. The proposed solution for labeling Web page pairs assumes that some domain specific knowledge is given and related Web pages must be connected by some link paths. The above assumptions restrict the applicability of the proposed solution. This therefore leaves much room for further research in homepage relationship mining.

1.2 Homepage Relationship Mining

In this paper, we denote a relationship r_k between concepts (or categories) c_s and c_t by $r_k(c_s, c_t)$, where c_s and c_t are the *source concept* and *target concept* respectively. For simplicity, only binary relationships will be considered and we shall leave the more complex relationships for future research.

Definition 1 Homepage Relationship Mining.
Given a relationship $r_k(c_s, c_t)$ and a set of homepages $\{h_1, h_2, \cdots, h_n\}$, homepage relationship mining *problem is to determine whether a pair of homepages $\langle h_s, h_t\rangle$ is an instance of the relationship $r_k(c_s, c_t)$ where $h_s \in c_s$ and $h_t \in c_t$.*

As mentioned in Section 1.1, homepage relationship mining and homepage mining are closely related. Our general approach is to assume that homepage mining is conducted before homepage relationship mining as the homepages assigned with concept labels will reduce the amount of candidate homepage pairs to be processed in homepage relationship mining. Having said this, the number of candidate pairs can still be very large. We also assume that all the homepages are from a single Web site as semantically related homepages are more likely to be from the same Web site given the autonomous nature of Web sites [13].

In this paper, we propose to tackle the homepage relationship mining problem using a binary classification model where the items to be classified are homepage pairs. This classification model essentially consists of three steps:

- *Candidate homepage pair generation*: This refers to constructing homepage pairs for classification using the labeled homepages.
- *Feature acquisition and representation*: This refers to extracting features from candidate homepage pairs and representing them in a form appropriate for the classification technique to be used.
- *Classifier construction*: Homepage pairs assigned with relationship labels by human expert(s) are used to train classifiers. Once learnt, classifiers can be applied on other unlabeled homepage pairs.

Though simple it may seem, there are some research challenges in the above classification approach to homepage relationship mining:

- Since each relationship instance involves a pair of homepages, there are many more content features that can potentially be defined. Suppose there could be n content features extracted from each homepage. There can be at least $2n$ content features for a homepage pair. If we consider other intermediate Web pages that connect the two homepages together, many more possible content features could be involved. Despite the large number of content features, almost none of them is relevant to homepage relationship mining. In fact, our experiments have shown that by simply concatenating content features from a pair of homepages, the relationship mining accuracy could be very poor. This suggests that a careful selection of features for homepage pairs is of utmost importance.
- The second challenge is that the number of candidate homepage pairs to be classified can be very large. Suppose there are 100 professor homepages and 200 course homepages. There will therefore be 20,000 possible candidate homepage pairs to be classified for the teach *(professor, course)* relationship. This causes significant overheads in the classification process.
- The third challenge is that the performance of the homepage relationship mining is closely affected by the quality of homepages and their concept labels. When there are errors in homepage mining, these errors will inevitably propagate to homepage relationship mining. It is therefore interesting to investigate ways to rectify the errors should they occur.

This paper focuses on the first two challenges as they are directly related to our proposed classification approach to homepage relationship mining. The third challenge concerns complex issues that will require a better understanding of the existing homepage mining and homepage relationship mining methods. We therefore leave it for future investigation.

1.3 Contributions

In this paper, we aim to develop *homepage relationship mining* methods that classify homepage pairs to pre-defined relationships. One can adopt the simple model that each homepage alone represents a concept instance and derive features for each homepage pair accordingly so as to construct relationship classifiers and conduct classification. This is also known as the **Web page based homepage relationship mining** (**PRM**) method.

Instead of treating a homepage as an individual concept instance, one can also view the other Web pages surrounding a homepage as part of the concept instance. A homepage with these neighborhood pages together form a **Web unit**. The Web unit abstraction for a concept instance captures more semantic information for deriving features of a homepage pair. In this research, we therefore develop the **Web unit based homepage relationship mining** (**URM**) method and conduct evaluation on it.

We state in the following the main contributions of this paper:

- *Definition of inter-homepage features*
  This paper introduces the notion of *inter-homepage features* to describe the background relations between a pair of homepages. The three kinds of inter-homepage features discussed in this paper are *navigation features*, *relative location features* and *common-item features*. We also develop the different sets of inter-homepage features for PRM and URM methods due to their distinctive models of concept instances.

- *Evaluation of homepage relationship mining methods*
  Using Support Vector Machines (SVM) as the base classifiers, we develop both the PRM and URM methods. We also experiment with different combinations of inter-homepage features. The experiments show that URM method outperforms PRM method. Furthermore, navigation features have been shown to give most discriminating outcome among the different types of inter-homepage features.

- *Zero-filter*
  Due to the potentially very large number of homepage pairs, we proposed a *zero-filter* to prune away homepage pairs that are unlikely to be related before they are examined by the classifiers. Such a filter can significantly reduced efforts in both training and classification without compromising much classification accuracy as shown in our experiments.

1.4 Paper Outline

The rest of the paper is organized as follows. In Section 2, we introduce the concept of Web unit and give an overview of the iterative Web unit mining algorithm. The three types of inter-homepage features and their associated supplementary features are discussed in Section 3. In Section 4, inter-homepage features are derived from homepage pairs considering their associated Web units. Experiments on WebKB dataset using both PRM and URM methods with different feature combinations are reported in Section 5. The related works are surveyed in Section 6. Finally we conclude this paper in Section 7.

h_1	http://⟨Website⟩/course/SC101/SC101.html
s_1	http://⟨Website⟩/course/SC101/lecture-programs.html
s_2	http://⟨Website⟩/course/SC101/instructors.html
s_3	http://⟨Website⟩/course/SC101/officehours.html
s_4	http://⟨Website⟩/course/SC101/exams/final.html
s_5	http://⟨Website⟩/course/SC101/exams/preliminary.html

Figure 2: Course Web unit example: SC101

2 Web Units of Homepages

Homepages are often created with links connecting to semantically related Web pages to provide supplementary information. For example, a professor homepage, say, *index.html*, may contain links to Web pages describing his research interests, curriculum vitae, education and professional experience etc.. Together with *index.html*, these pages form a complete professor concept instance, known as a professor *Web unit*.

Definition 2 Web Unit.
Given a Web site W, a Web unit u_i is a Web page or a set of Web pages from W that jointly provides information for a concept instance. A Web unit consists of exactly one homepage and zero or more support pages.

The homepage of u_i, also known as the key page, is denoted by $u_i.h$. The set of support page(s) of u_i are denoted by $u_i.s$. The homepage represents an entry point to reach all support pages of a Web unit[2]. In other words, the support pages of a Web unit are all reachable from the homepage through links. In our Web unit definition, a support page can only be part of one Web unit although it might be reachable from multiple homepages through links. Whether a page should be a support page of a Web unit depends on whether the page provides supplementary information for the logical entity described by the Web unit[3]. A Web unit example of course *SC101* is given in Figure 2. It consists of a homepage, h_1 (underlined in Figure 2), and 5 support pages, $\{s_1, s_2, s_3, s_4, s_5,\}$, providing additional information about SC101, including lecture programs, instructors, office hours and also information on examinations.

To illustrate Web units in reality, two Web graphs showing the linkage among Web pages from the Computer Science department of Cornell University in the WebKB dataset (see Section 5 for more detail on the dataset) are plotted in Figure 3. These Web pages are manually examined and are classified into

[2] Note that only few Web units having support pages that are not reachable from their homepages through links. Such cases are rare and can be eliminated.

[3] In some cases, the determination of support pages could be subjective.

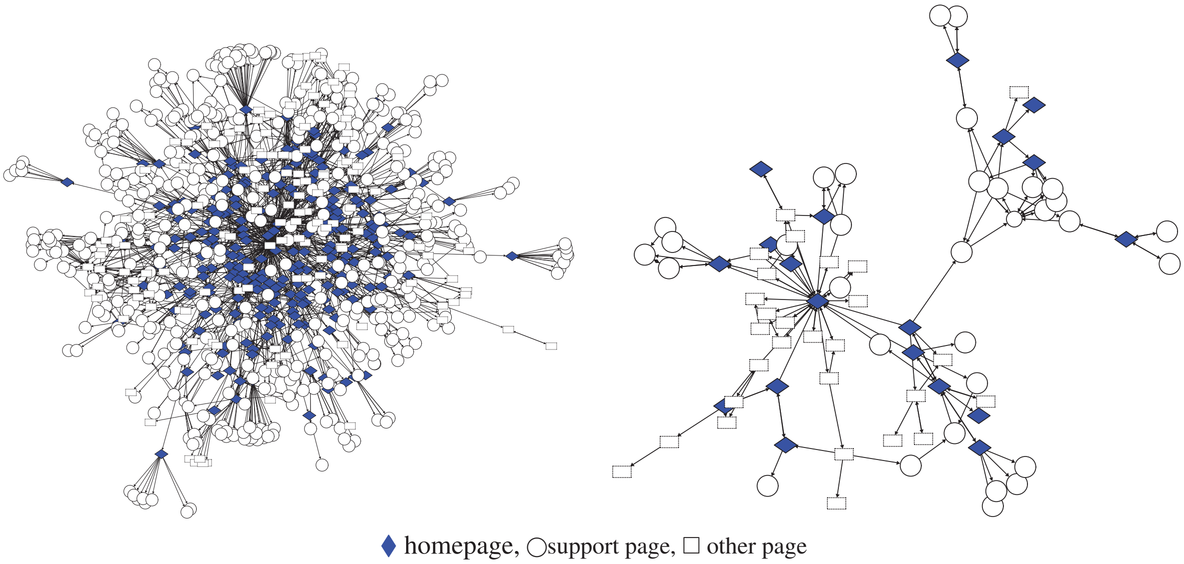

Figure 3: Web graphs plotted for Cornell Web pages in WebKB dataset. Graph(a) homepage 241, support page 458, other page 133, and link 1997; Graph(b) homepage 19, support page33, other page 28, and links 184.

homepages, support pages and other pages (i.e., does not belong to any Web units). Web graph (a) is produced by performing a breadth-first search starting from the Web site homepage[4]. Web graph (b) is a subgraph of (a) and is produced in a similar manner by ignoring the pages having more the 6 outgoing links (except the Web site homepage) during the breath-first search. This subgraph is less cluttered and allows us to observe the presence of Web units. The numbers of homepages, support pages and other pages are indicated below the two Web graphs. From the two Web graphs, the following observations can be made:

- Web units generally exist in Web sites; several Web units are clearly plotted around the outer border of the Web graph (a). Each Web unit has number of support pages ranging from zero to more than ten.
- Most homepages are plotted in the central area of the Web graph (a). There are more links among homepages than support pages, showing that homepages are likely to be the targets of links.
- There are hub pages (e.g., the page at the center of graph (a)) linking to a large number of homepages.
- Web units can be connected by links involving support pages and/or other pages besides direct links among their homepages. Such kinds of connections are clearly illustrated in Web graph (b).

The two Web graphs illustrate the existence of Web units in real Web site and also provide us the hints in determining the relationships among homepages using links.

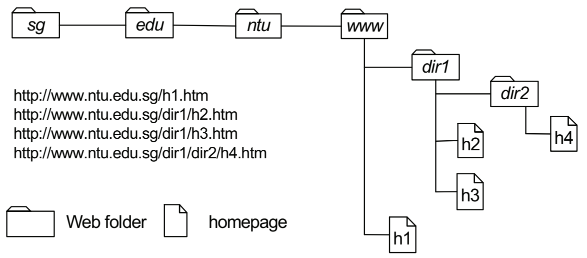
sg
edu
ntu
www
dir1
dir2
h4
h2
h3
h1
http://www.ntu.edu.sg/h1.htm
http://www.ntu.edu.sg/dir1/h2.htm
http://www.ntu.edu.sg/dir1/h3.htm
http://www.ntu.edu.sg/dir1/dir2/h4.htm
Web folder
homepage

Figure 4: Example Web directory derived from 4 URLs

2.1 Web Directory

To mine Web units, other than using links among pages, we also derive the Web site structure from the URLs of Web pages. Such a structure is known as *Web directory*. The URLs of homepages (and other Web pages) share the common format: *protocol type::*//*hostname [:port number] [/path] [filename]*. From the URLs, we can determine a set of *Web folders* from the elements of the *hostname* component separated by the delimiter "." and the elements of *path* component separated by the delimiter "/". Given a set of pages, a Web directory is therefore a tree consisting of Web folders and pages as nodes, and the inclusion relationship among them as edges. A Web directory derived from 4 example URLs is shown in Figure 4. Such Web directory structure is also utilized in our homepage relationship mining methods.

2.2 Web Unit Mining Algorithm for Homepage Mining

Given a collection of Web pages from a Web site, *Web unit mining* problem refers to the task of constructing Web units from

[4] URL: http://www.cs.cornell.edu/. It is also known as the *department* homepage for Cornell university in WebKB dataset

these Web pages and assigning them the appropriate concept labels. In this way, the homepages will also be discovered. Moreover, the homepages are assigned the concept label assigned to the Web units. To automatically perform Web unit mining, we assume that a set of perfectly labeled Web units are provided for training purpose. The training Web units may not come from the Web sites where the Web units are to be mined.

In our earlier work [18], we proposed the iterative Web unit mining method (iWUM) to mine Web units in an iterative manner. The algorithm draws ideas from both co-training and EM algorithms. As illustrated in Figure 5, there are two phases in iWUM, namely *Web fragment generation and classification* and *Web unit construction and classification*.

- *Web fragment generation and classification*
  A Web fragment is a Web page or a set of Web pages that can be considered as a potential Web unit or a portion of a Web unit. In the Web fragment generation phase, we take the input collection of Web pages and construct a Web directory representing the folder structure of the Web site. Once the Web directory is built, the Web folders likely to contain homepages of Web units are determined. From the selected Web folders, candidate homepages are identified and their Web fragments are generated. For example, given the six pages shown in Figure 2, we suppose that three Web fragments[5] can be generated: g_1={$\underline{h_1}$, s_1, s_2, s_3}, g_2={$\underline{s_4}$}, and g_3 ={$\underline{s_5}$}, where their homepages are underlined. The generated Web fragments are then classified and assigned appropriate labels by the classifiers constructed from the training Web units.

- *Web unit construction and classification*
  In this phase, Web units are constructed from the classified Web fragments based on some heuristic rules. Following our example, g_1, g_2, and g_3 could be merged together to form one course Web unit u_1={$\underline{h_1}$, s_1, s_2, s_3, s_4, s_5,}. With these constructed Web units, the information on how the Web site organizes these Web units are collected. For example, are Web units of the same concept located together under a common parent folder? Is there a hub page linking to all homepages of Web units of the same concept? Such information is used as features to construct Web unit classifiers that re-classify the Web units and assign them updated concept labels. This Web unit construction and classification process (enclosed in the dotted box in Figure 5) repeats itself until there are no changes or only very minor changes to the Web unit concept labels.

Detailed algorithm and experiment results of iWUM are given in [18]. Considering the fact that support pages and the homepage of a Web unit are closely-related, inter-homepage features derived from the support pages will also be informative in homepage relationship mining.

Table 1: PRM navigation features of homepage pair $\langle h_s, h_t\rangle$

id	connectivity type	*id*	connectivity type
n_1	$h_s \rightarrow h_t$	n_5	$h_s \leftarrow h_t$
n_2	$h_s \rightarrow p \rightarrow h_t$	n_6	$h_s \leftarrow p \leftarrow h_t$
n_3	$h_s \leftarrow p \rightarrow h_t$	n_7	$h_s \rightarrow p \leftarrow h_t$
n_4	$h_s \rightarrow p \rightarrow p \rightarrow h_t$	n_8	$h_s \leftarrow p \leftarrow p \leftarrow h_t$

3 Inter-homepage Features

Inter-homepage features are features carrying some information about the relationship between two homepages. They are remarkably different from the features (also called content features) that are typically used to assign concept labels to homepages. In other words, inter-homepage features are ones that bind a pair of related homepages together. In contrast, content features are extracted from individual homepages independently.

Definition 3 Inter-homepage Feature
Given a homepage pair $\langle h_s, h_t\rangle$, an inter-homepage feature $\gamma_{s,t}$ is a binary value indicating the existence of a kind of background relation between the two homepages.

In our research, we consider three kinds of background relations, namely, *navigation*, *relative location* and *common-item*. These inter-homepage features are known as *navigation features*, *relative location features*, and *common-item features* respectively. For each kind of background relation, *supplementary features* can also be derived for relationship mining. In the following, we define these three kinds of inter-homepage features and the supplementary features for a homepage pair.

- *Navigation Features*

  Links between Web pages have often been used to determine their relationships [3, 6]. Navigation features refer to the features derived from the linkage between a pair of homepages. This linkage may involve a direct link, a chain of links, or links with other connective structures.

  For example, Table 1 lists the 8 navigation features considered in our experiments. Each navigation feature represents a type of linkage between h_s and h_t. E.g., the direct-link navigation feature n_1 describes whether there exists a direct link from page h_s to page h_t, denoted by $h_s \rightarrow h_t$. The symbol → represents a directional link and p represents some Web page. The navigation feature is assigned the value of 1 if the corresponding linkage exists and 0 otherwise. The navigation feature n_2, $h_s \rightarrow p \rightarrow h_t$, describes whether h_t is reachable from h_s by 2 links.

- *Relative-location Features*

  Relative location features refer to the association between two homepages' locations in the *Web directory*. The Web directory represents the structure of a Web site constructed from the URLs of Web pages from the Web site (See Section 2.1).

[5] These three Web fragments are generated based on Web fragment generation algorithm and are used as an illustrative example. More details about the algorithm are available in [18].

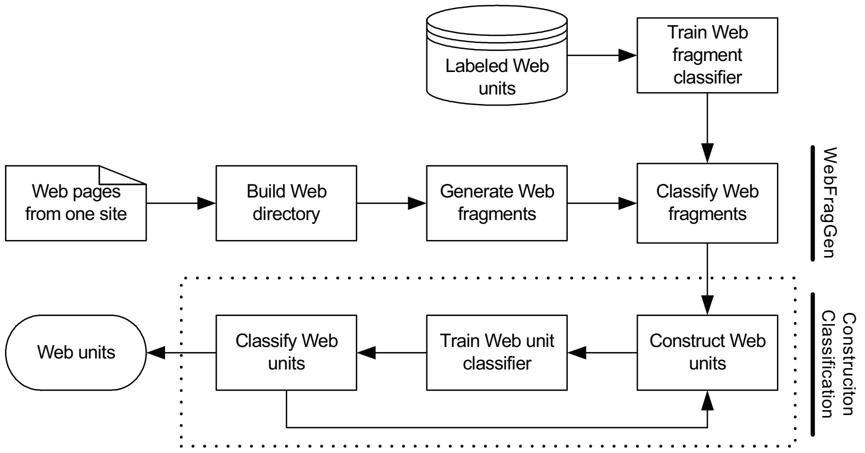

Figure 5: Iterative Web unit mining algorithm (iWUM)

In our experiments, we consider the locations of the two homepages in a candidate pair to be interesting when they are related by *parent-child*, *sibling* and *ancestor-descendent* in the Web directory. For example, in Figure 4, pages h_2 and h_3 are siblings in the Web directory; $h2$ and h_4 are related as parent-child; similarly, h_3 and $h4$ are also related as parent-child, and h_1 and h_4 are related as ancestor-descendent.

- *Common-item Features*

  Very often, common items appearing in a pair of homepages suggest some kind of relationship between them. They are some attributes or properties of concept instances cited by the homepages of related concept instances. Examples of common items include people names, email addresses, and telephone numbers. As the common items are attributes or properties of concept instances, the choice of which common items to use as features is therefore domain dependent.

  In our experiment, we used email address as the common item feature between a pair of homepages. One of the reasons is that in the university domain, any two homepages containing a common email address are likely to be related to the email address owner. The second reason is that email addresses can often be found within Web pages and they are easy to extract. Among the 4159 Web pages in the WebKB dataset used in our experiments, 2471 Web pages (or 59%) contain at least one email address. Altogether, we have extracted 5224 email addresses from these 2471 Web pages.

- *Supplementary Features*

  Although inter-homepage features suggest some relationship between pairs of homepages, they alone may not carry enough semantics for relationship mining. For example, links between homepages may be created for purposes other than relationship between them. Moreover, it may not be easy to determine the exact *type* of relationship between two homepages using link features only. To address these limitations, we introduce *supplementary features*. Supplementary features are additional features derived for some inter-homepage features and must be used together with their associated inter-homepage features.

  Supplementary features can be associated with all kinds of proposed inter-homepage features. Associated with the navigation features are the supplementary features derived from the linkage between a pair of homepages. For example, the anchor words associated with the links between a pair of homepages can be used as the navigation supplementary features. The anchor words normally carry some description about the target page of the associated link [10, 19]. Similarly, example supplementary features associated with common-item features are the words surrounding these common-items. Finally a pair of homepages with relative location features may use string tokens appearing in the folder/page names as supplementary features.

  In our experiments, we only employed the supplementary features associated with navigation features. These features are obtained from the anchor words associated with the links that directly connect two homepages, i.e., $h_s \rightarrow h_t$ and $h_s \leftarrow h_t$. The anchor words associated with links involving an intermediate page are not considered as they could be descriptions of the intermediate page and have little to do with the type of relationships held by a pair of homepages.

Note that in our discussions on inter-homepage features, we do not restrict to a specific representation (or a specific value) for each type of the inter-homepage features. The reason is that the inter-homepage features are used to represent background relations between homepages and the background relations in most cases are domain dependent. By not restricting the inter-homepage features, our homepage relationship mining approach can be easily applied to different domains.

4 Web Unit based Inter-homepage Features

In this section, we discuss how the inter-homepage features can be derived from the homepage pairs by considering their associated Web units.

- *Navigation Features*

  A Web unit is essentially a subgraph of Web pages. Different from the links connecting two pages as shown in Figure 6(a), a link in a Web unit can be either an *intra-unit link* connecting two Web pages from the same Web unit or an *inter-unit link* connecting a page from the Web unit to another page outside the Web unit. The inter-unit and intra-unit links are shown in Figure 6(b). Since the intra-links mainly capture the internal structure of a Web unit, and do not carry much information about the relationship(s) between homepages of two Web units, they are not used to derive navigation features in URM method as shown in Figure 6(c).

  Given a homepage pair $\langle h_s, h_t \rangle$, let u_s and u_t be the associated source and target Web units respectively. By distinguishing the homepage and support pages in a Web unit, an inter-unit link that directly connects u_s and u_t can be further categorized into one of the 8 types (n_1 - n_8) shown in the first column of Table 2 considering the direction of the inter-unit links. $u_s.h \rightarrow u_t.s$ holds if there exists at least one link from $u_s.h$ to any page in $u_t.s$. Similarly $u_s.s \rightarrow u_t.s$ holds if there exist at least one link from a page in $u_s.s$ to any page in $u_t.s$. When the two Web units u_s and u_t are connected through an intermediate page p, for example, $u_s \rightarrow p \rightarrow u_t$, p can be linked with either the homepage or any of the support pages in u_s and u_t. Hence 16 navigation features can be derived for u_s and u_t connected through an intermediate page (see n_9 - n_{16} in Table 2). Two Web units may be connected through two or more intermediate pages. We however believe that the information carried by that kind of inter-unit links is limited and we only consider the above 24 navigation features in our URM method.

- *Relative-location Features*

  For each Web unit, a Web directory can be derived from its Web pages. More specifically the Web directory derived from a Web unit is a sub-directory of the one derived from the Web site. Various relative-location features may be derived by comparing the locations of the two sub-directories and distinguishing the homepages and support pages. In our experiments, we simply use the relative-location features derived from the two homepages of the two Web units. In other words, relative-location features of URM method are the same as that of PRM method.

- *Common-item Features*

  We believe that if two Web pages contain common item(s), it may suggest that they are related. Similarly, we consider the common items appearing in two Web units. Since we distinguish the homepage and support pages in a Web unit, we define three types of common-item features in URM method.

  1. common items between $u_s.h$ and $u_t.h$,
  2. common items between either $u_s.h$ and $u_t.s$ or between $u_s.s$ and $u_t.h$,
  3. common items between $u_s.s$ and $u_t.s$.

  Note that common-item features are non-directional. Hence, we do not distinguish common-item features between $u_s.h$ and $u_t.s$ and the ones between $u_s.s$ and $u_t.h$.

- *Supplementary Features*

  For the easy comparison with PRM, supplementary features associated with navigation features are also considered in URM. In our experiments, these features refer to the anchor words associated with the links that directly connect two Web units and pointing to the homepages of Web units, i.e., the anchor words associated with $u_s.h \leftarrow u_t.h$, $u_s.h \leftarrow u_t.s$, $u_s.h \rightarrow u_t.h$, and $u_s.s \rightarrow u_t.h$. The anchor words associated with the links pointing to the support pages are not considered as these words are more for describing the target support pages.

5 Experimental Evaluation

We have enumerated several important kinds of inter-homage features. They represent the basic set of features that can be used for homepage relationship mining. We recognize that a comprehensive and detailed study of these features on datasets with different characteristics and sizes using different types of classification methods is an important but also very complex research task. Instead of trying out all the possibilities, our paper will begin with a few kinds of inter homepage features.

The homepage relationship mining experiments were conducted on WebKB and UnitSet datasets using SVM classifier. The WebKB dataset was used in our experiments mainly because (i) the Web pages and the relationships between Web pages have been labeled and (ii) the dataset is well known and has been used by previous works on similar research problems. Using WebKB will allow us to compare our work with them. SVM classifiers were used because of its good performance in text/Web page classification [8, 11, 19]. In our experiments, we used SVM^{light} package implemented by Joachims. Default parameters were used and instance pairs with classification score of greater than 0 were accepted as positive pairs. To avoid the handling of large number of negative relationship instances, we propose a *zero-filter* to discard homepage pairs in which the two homepages are not related to each other by any of the inter-homepage features.

5.1 Dataset

WebKB dataset contains Web pages collected from Computer Science departments of four universities. There are in to-

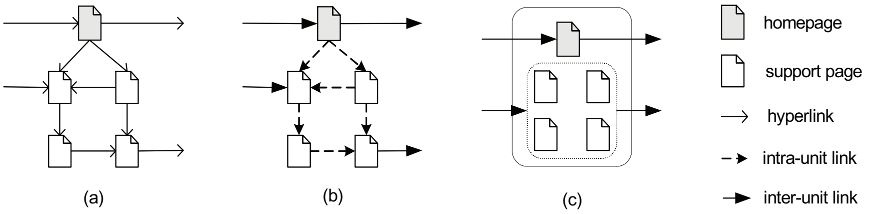

Figure 6: Difference between link, inter-unit link and intra-unit link

Table 2: URM navigation features for homepage pair $\langle u_s.h, u_t.h \rangle$

id	connectivity type	*id*	connectivity type	*id*	connectivity type
n_1	$u_s.h \rightarrow u_t.h$	n_9	$u_s.h \rightarrow p \rightarrow u_t.h$	n_{17}	$u_s.h \rightarrow p \leftarrow u_t.h$
n_2	$u_s.h \rightarrow u_t.s$	n_{10}	$u_s.h \rightarrow p \rightarrow u_t.s$	n_{18}	$u_s.h \rightarrow p \leftarrow u_t.s$
n_3	$u_s.s \rightarrow u_t.h$	n_{11}	$u_s.s \rightarrow p \rightarrow u_t.h$	n_{19}	$u_s.s \rightarrow p \leftarrow u_t.h$
n_4	$u_s.s \rightarrow u_t.s$	n_{12}	$u_s.s \rightarrow p \rightarrow u_t.s$	n_{20}	$u_s.s \rightarrow p \leftarrow u_t.s$
n_5	$u_s.h \leftarrow u_t.h$	n_{13}	$u_s.h \leftarrow p \leftarrow u_t.h$	n_{21}	$u_s.h \leftarrow p \rightarrow u_t.h$
n_6	$u_s.h \leftarrow u_t.s$	n_{14}	$u_s.h \leftarrow p \leftarrow u_t.s$	n_{22}	$u_s.h \leftarrow p \rightarrow u_t.s$
n_7	$u_s.s \leftarrow u_t.h$	n_{15}	$u_s.s \leftarrow p \leftarrow u_t.h$	n_{23}	$u_s.s \leftarrow p \rightarrow u_t.h$
n_8	$u_s.s \leftarrow u_t.s$	n_{16}	$u_s.s \leftarrow p \leftarrow u_t.s$	n_{24}	$u_s.s \leftarrow p \rightarrow u_t.s$

Table 3: Relationship instance distribution

University	department-of		instructor-of		member-of	
	Pos	Neg	Pos	Neg	Pos	Neg
Cornell	183	0	32	7654	66	3594
Texas	197	0	42	7444	89	3851
Washington	161	6	65	12293	135	3372
Wisconsin	207	3	112	17108	102	5148
Total	748	9	251	44499	392	15965

tal 4159 pages in the dataset. Among them, the homepages were manually classified into 6 categories, namely, *department*, *project*, *course*, *student*, *staff*, and *faculty*. The remaining pages were assigned to *other* category. Three relationships, namely, department-of(*person*, *department*), member-of(*person*, *project*) and instructor-of(*person*, *course*), have been manually labeled. Note that the *person* category is virtual as homepages belonging to *student*, *faculty* and *staff* are also considered as *person* pages (see Figure 1).

A homepage pair $\langle h_s, h_t \rangle$ where $h_s \in c_s$ and $h_t \in c_t$ is known as a positive relationship instance of relationship $r_k(c_s, c_t)$ if the relationship is manually labeled in WebKB dataset and negative relationship instance otherwise. The positive/negative relationship instances distribution for each university and each relationship in WebKB dataset is shown in Table 3. Note that in WebKB dataset, there is only one department homepage (computer science department) for each of the four university and almost all the person instances (i.e., student, staff and faculty) are from the department. As indicated in Table 3, there are much less negative examples for department-of relationship. By returning positive labels for all the department-of candidate relationship instances, nearly perfect results can be achieved. In other words, the department-of relationship does not add much information to our experiments. Therefore, in our discussion (including statistical significance test), department-of relationship is excluded. The experimental results for department-of are reported for completeness purpose only.

To conduct experiments on URM methods, we manually labeled Web units in the WebKB dataset. The labeling process is to find the support pages from *other* category for each homepage in the first six categories. We call the newly labeled dataset **UnitSet**. Few pages from the *other* category that cannot be labeled as support pages to any homepage as a Web unit were excluded from UnitSet. The number of Web units (denoted by $|u|$) and the total number of pages constituting these Web units (denoted by $|p|$) are reported in Table 4. As UnitSet is labeled based on WebKB dataset, the numbers of positive/negative relationship instances are the same.

5.2 Experimental Setup

All the three types of inter-homepage features and the supplementary features associated with navigation features discussed in Sections 3 and 4 were evaluated in our experiments. We denote navigation features by N, relative-location features by R, common-item features by E, and supplementary features by A respectively. Their dimensionality are presented in Table 5.

Table 5: Inter-homepage features and their dimensionality

Inter-homepage feature	PRM	URM
Navigation (N)	8	24
Relative location (R)	3	3
Common-item (E)	1	3
Supplementary navigation (A)	m	m

Performance of homepage relationship mining for a relationship r is measured by *precision* and *recall* denoted by Pr_r and

Table 4: Web unit distribution in UnitSet

Concept	*student*		*faculty*		*staff*		*course*		*project*		*department*	
University	$\|u\|$	$\|p\|$	$\|u\|$	$\|p\|$	$\|u\|$	$\|p\|$	$\|u\|$	$\|p\|$	$\|u\|$	$\|p\|$	$\|u\|$	$\|p\|$
Cornell	128	301	34	60	21	57	42	219	20	78	1	6
Texas	148	370	46	104	3	10	38	95	20	115	1	14
Washington	126	495	31	71	10	23	74	360	21	129	1	1
Wisconsin	156	416	42	83	12	27	82	413	25	90	1	8

Re_r respectively.

$$Pr_r = \frac{TP_r}{TP_r + FP_r} \tag{1}$$

$$Re_r = \frac{TP_r}{TP_r + FN_r} \tag{2}$$

where TP_r is the number of correctly labeled relationship instances and FP_r is the number of wrongly labeled among all relationship instances labeled by a relationship mining method; FN is the number of relationship instances that should be labeled to r but are mislabeled by the method.

$$F_{1_r} = \frac{2 \cdot Pr_r \cdot Re_r}{Pr_r + Re_r} \tag{3}$$

The F_1 measures computed from Pr and Re are also reported. The *macro* and *micro* averages of precision/recall are denoted by Pr^M/Re^M and Pr^μ/Re^μ respectively.

In our experiments, we used *leave-one-university-out* cross-validation. That is, in each run, relationship instances from three universities were used as training examples and the relationship instances from the fourth university were used as test data. The experimental results reported in Section 5 are the average of the four runs.

Table 6: Percentages of homepage pairs remained in PRM (%)

Features	department-of		instructor-of		member-of	
	Pos	Neg	Pos	Neg	Pos	Neg
NoFilter	100	100	100	100	100	100
N	97.3	77.8	96.8	24.2	95.9	30.6
R	98.7	100	56.2	35.5	38.8	31.8
E	1.5	0	52.9	0.1	15.6	0.5
NR	100	100	97.2	53.7	97.7	50.5
NE	97.3	77.8	98	24.3	96.9	30.7
RE	98.7	100	74.9	35.6	47.2	31.8
NRE	100	100	98.4	53.7	98.2	50.5

Table 7: Percentages of homepage pairs remained in URM (%)

Features	department-of		instructor-of		member-of	
	Pos	Neg	Pos	Neg	Pos	Neg
NoFilter	100	100	100	100	100	100
N	79.4	77.8	96	7.2	95.4	17.3
R	98.7	100	56.2	34.2	38.8	31.8
E	23.7	0	59.8	0.4	28.1	0.3
NR	100	100	96.4	39	97.2	42.4
NE	79.5	77.8	98	7.4	96.7	17.5
RE	98.7	100	80.5	34.5	55.1	31.9
NRE	100	100	98.4	39.2	97.9	42.5

5.3 Zero-Filter

As shown in Table 3, the number of negative relationship instances is much larger than the positive ones for the instructor-of and member-of relationships. To reduce the number of negative pairs, we propose the *zero-filter*. The assumption is that: if an homepage pair holds certain kind of relationship, the two homepages must be related by some inter-page features with none-zero values. If the feature vector obtained for an homepage pair is a *zero-vector*, i.e., all the inter-homepage features are zeros, it suggests that the two homepages are not related in any kind (at least for the background relations defined by the inter-homepage features). Such an homepage pair is expected not to belong to any relationship and is therefore filtered away.

To evaluate the usefulness of the zero-filter, we applied the zero-filter to both the positive and negative homepage pairs for each kind of relationships in the dataset. In PRM, the percentages of homepage pairs remained after applying the zero-filter for different combinations of inter-homepage features are shown in Table 6. It shows that *N* was the most important features that appeared in the positive pairs. Without *N*(i.e., *R*, *E*, and *RE*), large number of positive pairs were filtered away. On the other hand, the existence of a link (or links) between two homepages did not necessarily imply that they had some relationship as 30.6% of negative member-of homepage pairs were also connected by links. When all these three types of inter-homepage features were used, i.e., feature combination *NRE*, more than 98% of positive pairs were able to pass the zero-filter for all the three kinds of relationships. However, only about 53.7% and 50.5% of negative pairs remained after applying zero-filter for relationships instructor-of and member-of respectively. Note that the zero-filter did not filter away negative department-of homepage pairs as there are only 9 such pairs in the WebKB dataset.

Percentages of homepage pairs remained after applying the zero-filter for different combinations of inter-homepage features in URM are shown in Table 7. Compared with the results of PRM method, more negative pairs were filtered away while the percentages of positive pairs remain similar. In other words, inter-homepage features derived based on Web units were more effective in filtering away negative homepage pairs without affecting the positive pairs.

In our experiments, we found that the zero-filter while drastically reducing the time for training, did not affect the classi-

Table 8: Performance with/without zero-filter using *NREA* in URM

Zero-Filter	Macro Measures			Micro Measures		
	Pr^M	Re^M	F_1^M	Pr^μ	Re^μ	F_1^μ
Is used	0.914	0.855	0.879	0.945	0.916	0.930
Is not used	0.913	0.852	0.877	0.946	0.914	0.930

Table 9: Training time in CPU-seconds using *NREA* in URM

Zero-Filter	department-of	instructor-of	member-of
Is used	0.12	2.86	4.81
Is not used	0.13	6.77	6.63

fication accuracy of the SVM classifiers when feature combinations *NRE* and *NREA* were used (see Table 8 for the classification performance comparison using *NREA* feature combination). When the other feature combinations are used, applying zero-filter resulted in slightly degradation in recall value as some of the positive homepage pairs were filtered away. Applying zero-filter reduced the training time of SVM classifiers. The training time in CPU-seconds[6] with and without zero-filter using *NREA* feature combination is given in Table 9. It shows that applying zero-filter can reduce the training time effectively especially when large number of homepage pairs were given, e.g., homepage pairs of instructor-of.

5.4 Performance of PRM Method

The performance of PRM method using different feature sets is reported in Tables 10 and 11. Table 10 reports the macro-averaged measures for each relationship over the four universities. Table 11 shows macro/micro-averaged measures.

We experimented all combinations of inter-homepage features. Based on the experimental results, the following conclusions can be drawn.

- Using N alone, the SVM classifiers delivered very high precision for all the three relationships, and fairly good recall for instructor-of and member-of relationships. Pr^M and Pr^μ for the three relationships were higher than 90% while Re^M and Re^μ exceeded 78% and 84% respectively.

- By adding the features R, a decrease of precision and an increase of recall were observed for instructor-of. Inclusion of R did not affect the performance of member-of relationship mining. On the whole, better F_1^M and F_1^μ were obtained for member-of relationship

- The addition of the feature E to N increased recall of instructor-of and member-of relationships while the precision dropped.

- The supplementary feature A had negative effect to both instructor-of and member-of.

Table 11: Macro/micro-averaged performance of PRM method

Features	Macro Measures			Micro Measures		
	Pr^M	Re^M	F_1^M	Pr^μ	Re^μ	F_1^μ
N	0.931	0.782	0.837	0.955	0.846	0.897
NR	0.929	0.801	0.848	0.953	0.865	0.907
NE	0.919	0.786	0.834	0.943	0.850	0.894
NRE	0.917	0.803	0.844	0.941	0.866	0.902
NRA	0.900	0.786	0.834	0.949	0.861	0.903
NREA	0.914	0.777	0.832	0.952	0.851	0.899

Table 13: Macro/micro-averaged performance of URM method

Features	Macro Measures			Micro Measures		
	Pr^M	Re^M	F_1^M	Pr^μ	Re^μ	F_1^μ
N	0.915	0.777	0.827	0.932	0.799	0.860
NR	0.914	0.853	0.870	0.938	0.912	0.925
NE	0.917	0.794	0.839	0.935	0.812	0.869
NRE	0.910	0.863	0.876	0.935	0.922	0.928
NRA	0.895	0.847	0.867	0.942	0.912	0.927
NREA	0.914	0.855	0.879	0.945	0.916	0.930

- Based on F_1 measure, the best performance for instructor-of was achieved using *NR*. Similar performances were delivered for member-of using feature combinations *N*, *NR*, *NE*, and *NRE*. In other words, compared with R and E, N is the dominant feature for member-of.

5.5 Performance of URM Method

The performance of URM method using different feature combinations is reported in Tables 12 and 13. Using N alone, fairly good precision for both instructor-of and member-of, and good recall for member-of were achieved. Similar to PRM approach, inclusion of additional features, i.e., R, E and A did not affect much on member-of for both precision and recall. For instructor-of, inclusion of additional features increased the recall and generally decreased the precision; only when *NE* was used, better precision for instructor-of were reported compared with N. In terms of F_1 measure, the best performance for instructor-of was reported when *NREA* was used. For member-of, *NREA* gave the best precision and comparable recall. Considering both F_1^M and F_1^μ, the feature combination *NREA* gave the best homepage relationship mining performance. Nevertheless, our paired t-test on F_1 values [7] reported in Table 14 does not show that *NREA* is significantly better than the other feature combinations since the p-values of NREA versus other combinations are greater than 0.05. In other words, the navigation features are the dominant features for homepage relationship mining.

Compared with results of PRM method in Section 5.4, poorer precision but better recall and F_1 values were observed for most of the feature combinations in URM. Based on sta-

[6] The training time is reported by SVM^{light} on the computer with following configuration: 2GHz CPU, 1GB RAM, Microsoft® Windows® 2000 Professional Edition.

[7] The F_1 values for each feature combination were taken from the 8 runs when instances from each of the two relationships (instructor-of and member-of) from each of the four university was used as the test data in leave-one-university-out cross-validation (see Section 5.2).

Table 10: Homepage relationship mining performance of PRM method

Methods	department-of			instructor-of			member-of		
Features	*Pr*	*Re*	F_1	*Pr*	*Re*	F_1	*Pr*	*Re*	F_1
N	0.989	0.973	0.981	0.911	0.684	0.756	0.894	0.689	0.775
NR	0.987	1.000	0.994	0.906	0.715	0.777	0.894	0.689	0.775
NE	0.989	0.973	0.981	0.874	0.688	0.744	0.892	0.697	0.779
NRE	0.987	1.000	0.994	0.869	0.715	0.762	0.895	0.695	0.778
NRA	0.987	1.000	0.994	0.815	0.687	0.743	0.899	0.671	0.765
NREA	0.987	1.000	0.994	0.859	0.656	0.738	0.896	0.675	0.766

Table 12: Homepage relationship mining performances of URM method

	department-of			instructor-of			member-of		
Features	*Pr*	*Re*	F_1	*Pr*	*Re*	F_1	*Pr*	*Re*	F_1
N	0.988	0.796	0.875	0.879	0.651	0.724	0.879	0.884	0.881
NR	0.987	1.000	0.994	0.877	0.673	0.737	0.879	0.884	0.881
NE	0.988	0.797	0.876	0.884	0.695	0.759	0.879	0.890	0.883
NRE	0.987	1.000	0.994	0.864	0.698	0.750	0.879	0.890	0.883
NRA	0.987	1.000	0.994	0.797	0.675	0.726	0.901	0.866	0.882
NREA	0.987	1.000	0.994	0.850	0.701	0.761	0.903	0.865	0.882

Table 14: p-values for paired t-test on F_1 values

Features	N	NR	NE	NRE	NRA
NR	0.1551	–	–	–	–
NE	0.3114	0.4558	–	–	–
NRE	0.3647	0.5490	0.3626	–	–
NRA	0.9147	0.8145	0.1818	0.3779	–
NREA	0.4701	0.5903	0.9772	0.6988	0.3959

Table 15: Statistical significance test URM vs PRM

Values	*Pr*	*Re*	F_1
t statistic	-2.961	2.914	2.634
p-value	0.0129	0.0127	0.0233

tistical significance test[8] shown in Table 15, URM performed significantly better than PRM in recall but worse in precision. In terms of the combined measure, F_1, URM performed significantly better as the p-values is less than 0.05. We therefore argue that inter-homepage features derived based on Web units are more beneficial than the ones derived based on Web page in homepage relationship mining. This can be attributed mainly by two reasons:

- *Richer content in Web units*
  One of the major differences between Web unit and Web page is that Web unit contains support pages. The support pages in Web units may include additional information for homepage relationship mining. In other words, some background relation characterizing the relationship between two homepages may be found from their support pages. For example, a link between $u_s.s$ and $u_t.s$ may indicate that $u_s.h$ and $u_t.h$ are related. However, in PRM, these kinds of background relations cannot be captured. Moreover, more links between two Web units also means that more anchor words can be used as features. This could explain why the inclusion of *A* in URM had positive effect but not in PRM. Support pages in Web units also increase the chance of having common-items, like email addresses between a pair of homepages, increasing the chance of identifying background relations between homepages.

- *Distinction between intra-unit links and inter-unit links*
  We have shown that navigation features are the dominant inter-homepage feature in homepage relationship mining. Moreover, as shown in Figure 6, distinguishing the navigation features based on intra-unit and inter-unit links helps much in homepage relationship mining. Suppose p_1, p_2, p_3 are all support pages of a Web unit u_s, i.e., $p_1 \in u_s.s$, $p_2 \in u_s.s$, and $p_3 \in u_s.s$. As all the support pages of a Web unit is reachable from the homepage, $u_s.s \to u_t.h$ can be in any one of the following forms: (1) $u_s.h \to p_1 \to u_t.h$, (2) $u_s.h \to p_1 \to p_2 \to u_t.h$, (3) $u_s.h \to p_1 \to p_2 \to p_3 \to u_t.h$. PRM treats each of these connectivity types separately making it difficult to classify the homepage pairs. In other words, distinguish of intra-unit link and inter-unit link gives more space in representing connectivity between two homepages.

5.6 Performance Based on iWUM Results

From the experiments above, we knew that URM delivered best performance using feature combination *NREA*. We further applied the homepage relationship mining with the feature combination *NREA* on the Web units obtained from iWUM where the homepages and their corresponding Web units are not perfectly labeled.

In our earlier work [18], Web units of four concepts were mined using iWUM; the four concepts are: *course*, *student*, *faculty* and *project*. Performance of iWUM measured using

[8] Paired t-test were conducted based on performance measures for instructor-of and member-of on all feature combinations.

Table 16: iWUM mining results

Concept	student		faculty		course		project	
University	*Pr*	*Re*	*Pr*	*Re*	*Pr*	*Re*	*Pr*	*Re*
Cornell	0.963	0.812	0.861	0.912	0.917	0.786	0.600	0.300
Washington	0.979	0.932	0.973	0.783	0.568	0.553	0.000	0.000
Wisconsin	0.809	0.873	0.920	0.742	0.875	0.662	0.000	0.000
Texas	0.883	0.923	0.923	0.571	0.687	0.695	0.800	0.480

precision and recall are reported in Table 16. iWUM was able to mine *student* Web units with high precision and recall. For *faculty* Web units, high precision and relatively low recall were observed. iWUM was fine for *course* Web units but poor for *project* Web units. Web units of *department* and *staff* were not mined due to the lack of training examples. As there is only one department instance for each university, we added the four department instances to the mined Web units.

Results of URM based on iWUM mined Web units are reported in Table 17. For all the four universities, the method was able to achieve very good performance in terms of precision and recall on department-of due to good Web unit mining accuracy for the *student* and *faculty* concepts. For instructor-of, the macro-averaged precision was 0.77 and is acceptable with respect to that of URM method on perfectly labeled Web units, i.e., 0.85. However, macro-averaged recall was just slightly above 0.4 due to the poor recall of *course* concepts. Since iWUM performed not very well for *project* concept, our Web unit relationship mining method also delivered poor result for the member-of relationship. In particular, no *project* instance was correctly mined for *Texas* and *Washington*. In summary, the performance of homepage relationship mining was quite dependent on that of Web unit mining. The former can achieve high precision and recall values only if Web units can be mined with a high accuracy.

6 Related Work

Two subtasks in homepage mining are: (i) to identify the homepages and (ii) to assign the homepages appropriate concept labels. The task of identifying homepages is closely-related to the task of *homepage finding* in TREC with noticeable differences discussed in Section 1.1. Studies in homepage finding benefits homepage identifying task by listing out the features that can be used to determine whether a Web page is likely to be a homepage. URL-type and file name conversion are the two examples of such features [14, 20, 21]. There are four URL-types: *root*, *sub-root*, *path* and *file*. As reported in [14], the Web pages having URLs of the first three types constitute less than 8% of the WT10g dataset but contribute more than 94% of the homepages. It was also found that Web pages with file names containing ‘welcome’ and ‘home’ are likely to be homepages. These features have been used to derive heuristics for identifying homepages in Web unit mining.

The task of assigning homepage concept labels is generally known as a Web page classification task and has been heavily studied. Various features have been used in Web page classification including words in the content of the Web page [7, 22], words associated with the links pointing to the page [9, 19, 22] and the neighbor Web page concept labels [4, 15, 16]. In Web page classification, various classifiers have been experimented including Naïve Bayes, *k*-NN, Support Vector Machines (SVM) and rule-based classifiers [22]. Among them, SVM classifier is one of the best performing classifiers reported.

To the best of our knowledge, homepage relationship mining problem has been studied by the CMU Text Learning group [5]. In their research, each concept instance is assumed to be a single Web page. Various Foil-like learning algorithms were proposed to derive relationship classification rules and they were evaluated using the WebKB dataset. All these algorithms use relational representation of the Web data. Among these algorithms, the best performance results in terms of F_1 measure were achieved by Path-Foil-Pilfs [5]. The Path-Foil-Pilfs method is based on the assumption that any relationship between Web pages is represented by some *link path*. The idea is to consider the Web as a graph containing nodes as concept instances and edges as some relationship instances. By applying Foil’s hill-climbing search on the graph and some background knowledge (background relations), the method derives the classification rules accordingly. The background relations in Path-Foil-Pilfs consist of: *class* (*page*), *link-to* (*link, page, page*), *has-word* (*link*), *all-words-capitalized* (*link*), *has-alphanumeric-words* (*link*), *has-neighborhood-words* (*link*). Each of the background relations is a set of predicates. For example, *has-word* (*link*) is a set of predicates each indicating the words occurring on a given link. There is one predicate for each word in the vocabulary and each instance indicates an occurrence of the word in the given link. The Path-Foil-Pilfs method derives these predicates from links and anchor text only.

Note that the results reproduced in Table 18 are provided for reference purpose only and should not be directly compared with ours as the two homepage relationship mining problems are quite different. In our research, we assume that all the homepages (or Web units) have been labeled with concept labels. Therefore, we consider a candidate relationship instance $r_k(h_s, h_t)$ to be valid only if h_s and h_t are two homepages and $h_s \in c_s$ and $h_t \in c_t$ where c_s and c_t are source and target concepts respectively. In [5], the set of homepages is not given and therefore candidate relationship instances were generated from any two distinct Web pages from the dataset. As a result, large number of candidate instances were generated, e.g.,

Table 17: Performance of URM method on iWUM mined Web units

	department-of			instructor-of			member-of		
University	*Pr*	*Re*	F_1	*Pr*	*Re*	F_1	*Pr*	*Re*	F_1
Cornell	0.986	0.770	0.865	0.800	0.250	0.381	0.323	0.303	0.312
Texas	0.989	0.893	0.939	0.731	0.452	0.559	0.000	0.000	0.000
Washington	0.863	0.863	0.863	0.737	0.438	0.549	0.000	0.000	0.000
Wisconsin	0.968	0.884	0.924	0.812	0.500	0.619	0.477	0.618	0.538
MacroAve	0.952	0.853	0.898	0.770	0.410	0.527	0.200	0.230	0.213
Overall	Pr^M:0.640 Re^M:0.498 F_1^M:0.546 Pr^μ:0.826 Re^μ:0.599 F_1^μ:0.695								

Table 18: Performance for Path-Foil-Pilfs method

department-of			instructor-of			member-of		
Pr	*Re*	F_1	*Pr*	*Re*	F_1	*Pr*	*Re*	F_1
0.984	0.758	0.857	0.869	0.606	0.714	0.810	0.554	0.658

more 300,000 negative relationship instances reported for each relationship [5]. In their experiments, the background relation *class*(*page*) is a set of binary predicates indicating the concept label assigned to each web page by some learning algorithm and these binary predicates were used as features in relationship mining.

7 Conclusions

Homepage relationship mining is an important problem. By classifying pairs of homepages into pre-defined relationships, we facilitate the construction of the semantic Web over the existing Web pages and Web sites. The main contributions of this paper are to (i) identify the various types of inter-homepage features that can be extracted from a homepage pair, and (ii) compare the homepage relationship mining performance using Web page based approach and Web unit based approach. Experiments on the WebKB and UnitSet datasets using SVM showed that the use of navigation, relative location, common-item features and supplementary features could gave very good homepage relationship mining results. Our experiments also showed that Web unit based approach outperformed Web page based approach. We also proposed a zero-filter to reduce the training and classification efforts.

As part of our future work, we will look into some extensions of homepage relationship mining research. One of them is the investigation of using Web blocks to derive inter-homepage features instead of Web pages or Web units. Web blocks refer to segments of Web pages that represent some logical units of information. In [2], links from Web blocks were studied to determine importance of Web pages and to improve search accuracy. As present, we assume that homepage mining is performed before homepage relationship mining. It is interesting to investigate if homepage relationship mining can be performed together with (or before) homepage mining. In this work, our homepage relationship mining methods have been experimented in university domain because of the availability of dataset. We would like to study the performance of our methods in other domains once suitable datasets are available.

References

[1] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2), Fall 2002.

[2] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Block-based web search. In M. Sanderson, K. Jarvelin, J. Allan, and P. Bruza, editors, *Proc. of the 27th ACM SIGIR*, pages 456–463, Sheffield, United Kingdom, 2004. ACM Press.

[3] K. S. Candan and W.-S. Li. Reasoning for web document associations and its applications in site map construction. *Data & Knowledge Eng.*, 43(2):121 – 150, Nov. 2002.

[4] S. Chakrabarti, B. E. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In L. Haas, P. Drew, A. Tiwary, and M. Franklin, editors, *Proc. of ACM SIGMOD*, pages 307–318, Seattle, USA, June 1998. ACM Press.

[5] M. Craven and S. Slattery. Relational learning with statistical predicate invention: Better models for hypertext. *Machine Learning*, 43(1-2):97–119, 2001.

[6] J. Dean and M. R. Henzinger. Finding related pages in the world wide web. In *Proc. of 8th WWW*, pages 389–401, Toronto, Canada, May 1999. Elsevier North-Holland, Inc.

[7] S. T. Dumais and H. Chen. Hierarchical classification of Web content. In E. Yannakoudakis, N. J. Belkin, M.-K. Leong, and P. Ingwersen, editors, *Proc. of 23rd ACM SIGIR*, pages 256–263, Athens, Greece, July 2000. ACM Press.

[8] S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proc. of 7th ACM CIKM*, pages 148–155, Bethesda, Maryland, Nov. 1998. ACM Press.

[9] E. Glover, K. Tsioutsiouliklis, S. Lawrence, D. Pennock, and G. Flake. Using web structure for classifying and describing web pages. In D. Lassner, D. D. Roure, and

A. Iyengar, editors, *Proc. of 11th WWW*, pages 562–569, Honolulu, Hawaii, May 2002. ACM Press.

[10] D. Hawking and N. Craswell. Overview of the TREC-2001 web track. In E. M. Voorhees and D. K. Harman, editors, *Proc. of TREC*, Maryland, November 2001. NIST Special Publication 500-250. Available online: http://trec.nist.gov/.

[11] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nedellec and C. Rouveirol, editors, *Proc. of 10th ECML*, pages 137–142, Chemnitz, Germany, Apr. 1998. Springer.

[12] M.-Y. Kan. Web page categorization without the web page. In S. I. Feldman, M. Uretsky, M. Najork, and C. E. Wills, editors, *Proc. of 13th WWW*, pages 262 – 263, New York, NY, USA, May 2004. ACM Press.

[13] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[14] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In K. Jarvelin, M. Beaulieu, R. Baeza-Yates, and S. H. Myaeng, editors, *Proc. of 25th ACM SIGIR*, Tampere, Finland, Aug. 2002. ACM Press.

[15] Q. Lu and L. Getoor. Link-based text classification. In T. Fawcett and N. Mishra, editors, *Proc. of 20th ICML*, pages 496–503, Washington, DC, Aug. 2003. AAAI Press.

[16] H.-J. Oh, S. H. Myaeng, and M.-H. Lee. A practical hypertext categorization method using links and incrementally available class information. In E. Yannakoudakis, N. J. Belkin, M.-K. Leong, and P. Ingwersen, editors, *Proc. of 23rd ACM SIGIR*, pages 264–271, Athens, Greece, 2000. ACM Press.

[17] E. O'Neill, B. F. Lavoie, and R. Bennett. Trends in the evolution of the public web, 1998-2002. *D-Lib Magazine*, 9(4), April 2003. Aavailable online: http://www.dlib.org/dlib/april03/lavoie/04lavoie.html.

[18] A. Sun and E.-P. Lim. Web unit mining – finding and classifying subgraphs of web pages. In D. Kraft, O. Frieder, J. Hammer, S. Qureshi, and L. Seligman, editors, *Proc. of 12th ACM CIKM*, pages 108–115, New Orleans, LA, USA, Nov. 2003. ACM Press.

[19] A. Sun, E.-P. Lim, and W.-K. Ng. Web classification using support vector machine. In R. H.-L. Chiang and E.-P. Lim, editors, *Proc. of 4th ACM WIDM held in conj. with CIKM*, Virginia, USA, Nov. 2002. ACM Press.

[20] T. Upstill, N. Craswell, and D. Hawking. Query-independent evidence in home page finding. *ACM Trans. Inf. Syst.*, 21(3):286–313, 2003.

[21] T. Westerveld, D. Hiemstra, and W. Kraaij. Retrieving web pages using content, links, urls and anchors. In E. M. Voorhees and D. K. Harman, editors, *Proc. of 10th TREC*, Maryland, 2001. NIST Special Publication 500-250. Avaliable online: http://trec.nist.gov/.

[22] Y. Yang, S. Slattery, and R. Ghani. A study of approaches to hypertext categorization. *J. of Intelligent Info. Sys.*, 18(2-3):219–241, 2002.