氏　　名	Wang Yuxin
学位（専攻分野）	博士（情報学）
学位記番号	総研大甲第1000号
学位授与の日付	平成１８年９月２９日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第６条第１項該当
学位論文題目	Study on Building a High-Quality Homepage Collection from the Web Considering Page Group Structures
論文審査委員	主　査　教授　大山　敬三 教授　武田　英明 教授　高須　淳宏 教授　相澤　彰子 教授　神門　典子 教授　安達　　淳（国立情報学研究所）

論文内容の要旨

This thesis is devoted to investigate the method for building a high-quality homepage collection from the web efficiently by considering the page group structures. We mainly investigate in researchers' homepages and homepages of other categories partly.

A web page collection with a guaranteed high quality (i.e., recall and precision) is required for implementing high quality web-based information services. However, to build such a collection demands a large amount of human work because of diversity, vastness and sparseness of the web pages. Even though many researchers have investigated search and classification of web pages, etc., most of them are of best-effort type and pay no attention to quality assurance. Thus, we are investigating a method to build a homepage collection efficiently with assuring a given high quality, with the expectation that the investigated method will be applicable to the collection of various categories of homepages.

This thesis consists of seven chapters. Chapter 1 gives the introduction, and Chapter 2 presents the related works. Chapter 3 describes the objectives, the overall performance goal of the investigated system, and the scheme of the system. Chapters 4 and 5 discuss in detail the two processing steps of the method which are used for realizing the investigated system respectively. Chapter 6 further discusses the method for reducing the processing cost of the system. Finally, Chapter 7 concludes the research and discusses the future work.

In Chapter 3, taking into account the enormous size of the real web, a two-step-processing method is introduced at first, i.e., rough filtering and accurate classification. The former is for narrowing down the candidate page amount fast enough with required high recall. The latter is for accurately classifying the candidate pages into three classes: assured positive, assured negative, and uncertain, with assuring required recall and precision.

We present in detail the configuration, the experiments, and the evaluation of the rough filtering in Chapter 4. The rough filtering is a method for gathering researchers' homepages (or entry pages) by applying our original simple and effective page group models for exploiting the mutual relations between the structure and the content of a page group. It aims at narrowing down the candidates with a very high recall. First, 12 property-based keyword lists that correspond to researchers' common properties are created and are assigned as either organization-related or non-organization-related. Next, four page group models (PGMs) are introduced taking into consideration the structure in an individual logical page group; PGM_Od: the out-linked pages in the same and lower directories, PGM_Ou: the out-linked pages in the upper directories, PGM_I: the in-linked pages in the same and the upper directories, and PGM_U: the site top and the directory entry pages in the same and the upper directories.

Based on the PGMs, the keywords are propagated to a potential entry page from its surrounding pages, composing a virtual entry page. Finally, the virtual entry pages that scored at least a threshold value are selected. Since applying PGMs generally causes a lot of noises, we introduced four modified PGMs with two original techniques, i.e., the keywords are propagated based on PGM_Od only when the number of out-linked pages in the same and lower directories is less than a threshold, and based on the other PGMs, only the organization-related keywords are propagated. The four modified PGMs are used in combination in order to utilize as many informative keywords as possible from the surrounding pages.

The effectiveness of the method is shown by comparing it to a single-page-based method through experiments using a 100GB web data set and a manually created sample data set. The experiment results show that the output pages from the rough filtering is less than 23% of 100GB data set by using the four modified PGMs in combination, under a condition that the recall is more than 98%. Another experiment using a 1.36TB web data set with the same rough filtering configuration shows that the output pages is less than 15% of the corpus.

In Chapter 5, we present in detail the configuration, the experiments, and the evaluation of the accurate classification method. Using two types of component classifiers (a recall-assured classifier and a precision-assured classifier) in combination, we construct a three-way classifier that inputs the candidate pages output by the rough filtering and classifies them to three classes: assured positive, assured negative, and uncertain. Here, the assured positive output assures the precision and the assured positive and uncertain output assures the recall, and hence only the uncertain output should be manually assessed in order to assure the quality of the web data collection.

We first devise a feature set for building the high performance component classifiers using support vector machine (SVM). We use textual features obtained from each page and its surrounding pages. The surrounding pages are grouped based on connection types (in-link, out-link, and directory entry) and relative URL hierarchy (same, upper, or lower in the directory hierarchy), then an independent feature subset is generated from each group. The feature subsets are further concatenated to compose a feature set for a classifier. We use two types of textual features (plain-text-based and tagged-text-based). The classifier using only the plain-text-based features in each page alone is used as the baseline. Various feature sets are tested in the experiment using manually prepared sample data and the classifiers are tuned by two methods, i.e., *offset*-based and *c-j-option*-based. The results show that the performance obtained by using *c-j-option*-based tuning method is statistically significant at 95% confidence level. The F-measures of the baseline and the top two performed classifiers are 83.26%, 88.65%, and 88.58%, and show that the proposed method is evidently effective.

In order to know the performances of the classifiers with the above mentioned feature sets for more general cases, we experimented with our method on Web->Kb data set, a commonly used test collection for the web page classification task. It contains seven categories and four of them, "course", "faculty", "project", and "student", are used for comparing the performance with prior works. The experiment results show that our method outperformed all the seven prior works based on macro-averaged F-measure and 10 out of 12 on per-category basis (F-measures of individual category are not available for 4 of the prior works). Therefore, we can conclude that our method performs fairly well and is applicable not only to the researchers' homepages but also to other categories and/or in other language.

By tuning the well performing classifiers independently, we then build a recall-assured classifier and a precision-assured classifier, and compose the three-way classifier using them in combination. We estimated the numbers of the pages to be manually assessed for the required precision/recall at 99.5%/98%, 99%/95%, and 98%/90%, using the output pages from a 100GB data set through the rough filtering. The results show that the manual assessment cost can be reduced down to 77.6%, 57.3%, and 51.8%, respectively, compared to the baseline. We did analysis on classification result examples and the result shows the effectiveness of the classifiers.

In Chapter 6, the cascaded structure of the recall-assured classifiers, used in combination with the rough filtering, is proposed for reducing the computer processing cost. The estimation on the numbers of pages requiring feature extraction in the accurate classification for 100GB and 1.36TB data sets shows that the computer processing cost can be reduced down to 27.5% and 18.3%, respectively.

In Chapter 7, we summarize our contributions. In over all, we presented a realistic framework for building a high-quality web page collection with two-step processes: the rough filtering and the accurate classification in order to reduce the processing cost. In the rough filtering, we contributed two original key techniques used in the modified PGMs to reduce irrelevant keywords to be propagated. One is to introduce a threshold on the out-linked page number in the same and lower directories, and the other is to introduce keyword list types and to propagate only the organization-related keyword lists from the upper directories. In the accurate classification, we contributed mainly in two aspects. One is our original method for exploiting features from the surrounding pages and concatenating the features independently to improve web page classification performance. The other is to use a recall-assured classifier and a precision-assured classifier in combination as a three-way classifier for reducing the amount of pages requiring manual assessment under the given quality constraints.

We also discuss the future works including, for the rough filtering, to find a more systematic way for modifying the property set and property-based keywords; for the accurate classification, to investigate the way to estimate the likelihood of the component pages and to incorporate them; and we will further utilize the information from the homepage collection for practical applications.

論文の審査結果の要旨

公開発表会において学位請求論文の内容に関する45分の発表と15分の質疑応答を英語で行った後，引き続き審査委員による論文審査を行った。公開発表会においては背景・目的，関連研究等も含めて本研究の内容について発表を行うとともに，質疑応答を行った。また，論文審査においては，予備審査以降に行った実験や分析の内容，及び予備審査において指摘された問題点・疑問点を中心に発表並びに質疑応答を行った。本論文は，品質保証型の情報サービス等の応用にも利用可能な高品質のWebページコレクションを構築するため，精度と再現率を保証しながら指定されたカテゴリのホームページをWebから収集するための方式を構築することを目的とするものである。コレクションの品質保証に必要となる人手判定を削減するとともに，膨大なWebデータを効率的に計算機処理することを目標としており，研究者ホームページを中心とし，その他の研究活動に関連したページをも対象として実験・分析を行っている。これらの目標の達成のため，本論文では，ホームページとその周辺ページからなるページグループの内容を，その構造に基づいて効果的に活用するためのモデル（ページグループモデル）を提案するとともに，それを実データへ適用のために“Rough Filtering”（RF）と“Accurate Classification”（AC）の二段階からなるシステム構成を提案している。RFにおいてはページグループモデルとそれに基づくキーワード伝播を用いて高効率かつ高再現率で候補ページを絞り込む手法を実装した。またACにおいてはページグループモデルを反映させた独自の素性集合を用いてWebページに対するSVMの分類性能を大幅に高めるとともに，精度保証付き分類器および再現率保証付き分類器を組み合わせて，自動分類による品質保証が不可能なWebページを正確に選択する手法を考案し，これらを組み合わせることにより人手判定対象を大幅に削減することを可能とした。

Webデータは大量かつ多様であるため，通常，Webデータを対象とした計算機処理によるデータ収集・分類等はベストエフォート型であるのに対し，本研究は品質保証手段を提供するという点で独創的かつ有用性の高い研究といえる。また，周辺ページの内容を利用するというアイデアは一般的であるにもかかわらずこれまで有効活用に成功した例はわずかであるのに対し，本研究では比較的簡潔な手法により安定的かつ顕著な性能改善を実現しており，汎用性の高い手法といえる。提案手法の有効性・信頼性については，実際のWebから収集されたデータセットを用いた正確かつ詳細な実験とその結果の分析により十分に示されている。これらの研究成果により，高品質のWebページコレクションを現実的なコストで構築する方式が示され，Webデータを活用した品質保証型の情報サービスの可能性が開けたと言える。

また，論文構成も適切であり，関連研究も十分な文献調査と独自の観点からの分析に基づき記述されている。

以上のように，本研究は独創性，有用性，有効性，信頼性とも十分に高いものであることから，本学位請求論文は学位授与に必要な基準に達しているものと認め，論文審査を合格と判定する。