様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 6 月 17 日現在

研究種目：基盤研究（B）
研究期間：2006～2008
課題番号：１８３０００３７
研究課題名（和文）大規模WWWデータからの情報資源構築のための高性能分類方式の研究
研究課題名（英文）Study on High Performance Classification Method for Constructing Information Resources from Large Scale WWW Data

研究代表者
大山　敬三（OYAMA KEIZO）
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：90177022

研究成果の概要：ウェブデータから情報資源を構築する際の省力化には，ウェブページの自動分類の精度を高める必要がある。本研究では，周辺ページの内容を有効に活用して分類性能を高めるため，ウェブサイト内のリンクとディレクトリ階層に表現された潜在的意味を活用する手法，及び分類に悪影響を与える周辺ページを除去する手法を開発し，実験により有効性を確認した。本手法により，人手による確認・判定作業を大幅に削減することが可能となった。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	2,300,000	690,000	2,990,000
2007 年度	2,900,000	870,000	3,770,000
2008 年度	2,600,000	780,000	3,380,000
年度			
年度			
総　計	7,800,000	2,340,000	10,140,000

研究分野：総合領域
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：Web ページ分類，テキスト分類，機械学習，周辺ページ，性能保証，判定コスト，情報資源，情報検索

１．研究開始当初の背景

日常生活や経済活動などのあらゆる場面で Web 情報の重要性が増すにつれ，Web からの情報収集技術に対する要求も強まるとともに多様化している。一方で Web データにおいては，情報の粒度や表現形式，内容，情報発信意図等の多様性の拡大，データ量の爆発，所望の情報の散在化など，情報収集の困難度は増大する一方である。これらに対処するためには Web 検索エンジンだけでは不十分であり，Web ページの自動分類技術の重要性が増している。

２．研究の目的

本研究では，様々な応用において利用可能な品質を持つ情報資源を構築することを目指して，指定されたカテゴリの Web ページコレクションを所与の高い再現率及び精度を満たしつつ効率的に収集するための分類方式を実現することを目的とする。

例えば研究情報ナビゲーションシステムのような応用を想定し，論文，研究者，プロジェクトといったエンティティ間の正確なリンケージを実現するためには，まずこれらに関する情報を様々な情報源から収集し，相互の対応関係を同定する必要がある。通常の

Web 検索や Web 情報抽出では良質の情報を収集することを目的とするのに対し，このような応用では，あるカテゴリに含まれる Web ページを網羅的に収集することが求められる（高再現率)。また，処理結果を直接人間が利用するだけであれば多少の誤りの混入は許容されるが，他のデータと密にリンケージを取る場合には誤りの混入は許されない（高精度)。

このため，Web ページコレクションの構築においては，所与の高い精度と再現率に対応した精度保証型と再現率保証型の各高性能分類器を作成し，Web ページがあるカテゴリに属するかどうかを「適合」,「不適合」,「未確定」に分類するための 3 段階分類器を構成し，未確定に分類されたものを人手判定することが現実的な解となる。一般的にはこのうち人手判定が最もコストが高いため，効率化のためには精度保証型および再現率保証型の分類器の性能を高めて未確定に分類される Web ページを可能な限り減らすことが中心課題となる。

ここで，Web 文書分類において高再現率を実現するためには，Web 文書の作成者やその所属機関が提供する，確実に存在する情報，すなわち，同一 Web サイト内にある周辺ページの内容を活用することが重要である。しかし，周辺ページには雑多な情報も多く含まれ，従来の手法ではこれが性能の低下を招く大きな要因となっていた。そこで本研究では，対象ページと周辺ページとの間の意味的関係を潜在的に表現すると考えられる形式的関係に着目し，それらを活用して分類性能を高度化する手法を開発することを目指す。特に精度保証型と再現率保証型の分類器に必要とされる高精度領域及び高再現率領域での性能向上を通じて，人手判定を要する未確定 Web ページの大幅な減少を目標とする。

３．研究の方法

従来研究においては，第三者による（Web サイトをまたがる）ハイパーリンク参照やアンカーテキストなどを活用して Web ページ分類の高性能化を実現する手法が多く提案されている。これらは認知度の高い Web ページには有効であるが，例えばできたての Web ページの分類には全く効果がないため，再現率の改善には限界がある。一方，同一 Web サイト内にある周辺ページを分類に活用する試みも多いが，明確な性能向上は得られておらず，多くの場合はむしろ性能低下を引き起こしている。

我々は，周辺ページの内容を活用した検索性能の向上という課題に対して，以下の二つの新しい手法によりアプローチする。(1) Web サイト内のリンク関係とディレクトリ階層に基づき周辺ページをグループ化し，ページ間の潜在的意味関係を活用する手法（周辺ページグループ手法）により，高再現率の領域を中心に高性能化を行う。(2) 周辺ページ中に含まれていて分類精度に悪影響を及ぼすノイズページを，リンクや内容などに関するページ間の関係に基づき除去する手法（ノイズページフィルタリング手法）により，高精度の領域を中心に高性能化を図る。

本研究では，分類アルゴリズムには文書分類での有効性が確認されている Support Vector Machine (SVM)を用い，以下の 2 つのデータセットを用いて実験を行った。(1) 約 1000 万件の日本語 Web ページコレクション NW100G-01 から抽出したサンプルに対して，研究者ホームページであるか否かを人手判定して作成した ResJ-2。(2) CMU の World Wide Knowledge Base プロジェクトが主要な大学の計算機科学科の Web サイトから収集し 7 カテゴリに分類した WebKB（カテゴリとしては course, faculty, project, student を使用)。

周辺ページグループ手法においては，Web ページテキストから内容語を抽出して特徴語とし，各周辺ページグループ中の特徴語の有無を素性として分類を行う。ページグループの構成においては接続タイプ（インリンク，アウトリンク，ディレクトリエントリ）とディレクトリ階層（上位ディレクトリ，同一ディレクトリ，下位ディレクトリ）との様々な組合せについて試行を行い，各部の周辺ページの有効性を確認するとともに最適な構成を探索する。

ノイズページフィルタリング手法においては，対象ページと周辺ページの間の関係を表現する各種の素性を用いて，対象ページ・周辺ページ対(ES-pair)を単位として分類を行う。訓練データとしては，新たな人手判定や外部情報を用いることなく，適合ページを対象ページとする ES-pair を適合データ，それ以外を不適合データと見なして SVM の学習を行う。ただし，擬似的な訓練データを用いているため，SVM の判別超平面のオフセットを調整してフィルタの最適化を行う。

なお，上記の手法に必要な処理は，対象ページ及び周辺ページの本文テキストやアンカーテキスト，URL，HTML 要素などを HTML 構造解析や形態素解析等のテキスト解析処理により分節化し，包含関係を抽出する程度の比較的軽いものであるが，大量の Web データ全てに対して行うことは容易ではない。そこで，テキスト解析処理を必要とせず，文字列マッチングにより再現率優先で候補ページを高速に収集する手法も併せて導入する。

４．研究成果

本システムの全体構成を図 1 に示す。Rough filtering は大量の Web データを効率よく処理するために，キーワードマッチング

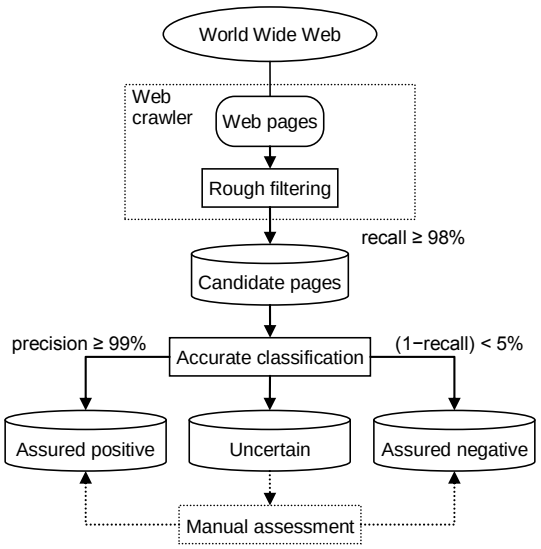

図１　分類システム全体構成

により再現率優先で適合の可能性のある Web ページを収集する。Accurate classification は本研究の中心となる３段階分類器である。

Rough filtering では，まず，概念的には図２に示す方法により周辺ページの内容語を集約し，文書ベクトルを作成する。次に，経験的に作成した１２種類のキーワードリストと照合し，含まれるキーワードの種類数をスコアとする。そして，訓練データより，要求された再現率を保証可能なスコアの閾値を求め，それ以上のスコアを持つ Web ページを候補として出力する。

Accurate classification には図３のような構成の３段階分類器を用いる。再現率保証

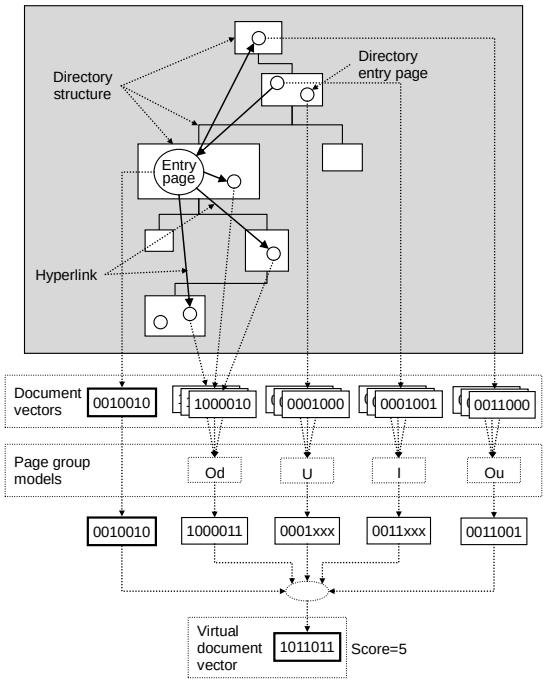

図２　Rough filtering 用文書ベクトル

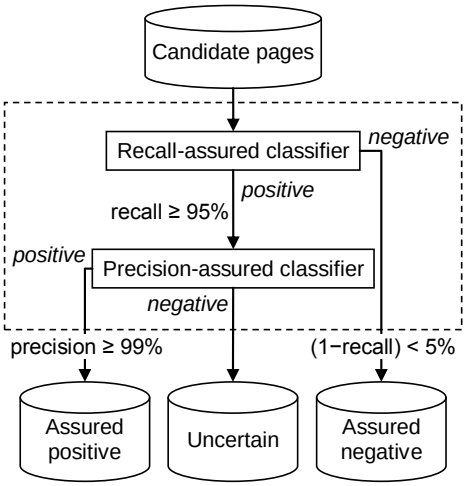

図３　３段階分類器の構成

型（Recall-assured）と精度保証型（Precision-assured）の各基本分類器は図４のような構成を取る。

周辺ページ分類器（Surrounding page classifier）ではノイズページフィルタリング手法を用いて，以下の分類に悪影響を及ぼす可能性のある周辺ページを除去する。対象ページと周辺ページとの間の関係を表現する素性として，リンク・URL 関連素性，アンカーテキスト・内容間関連素性，共通内容語関連素性，ページスタイル関連素性，URL・アンカーテキスト単語関連素性の５種類を用いて SVM による分類を行っている。SVM の判別超平面のオフセットは，予備実験の結果から，テストデータにおける再現率が 80%となるように調整した。

対象ページ分類器（Entry page classifier）では周辺ページ分類器を通過した周辺ページに対して周辺ページグループ手法を用いて素性を抽出し，対象ページに対する高性能分類を実現する。概念的には図５に示す方法により文書ベクトル（素性集合）

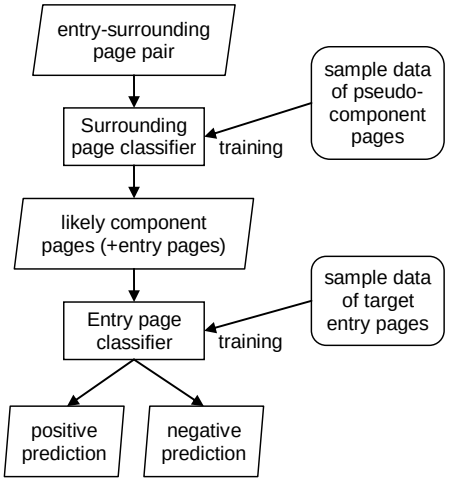

図４　基本分類器の構成

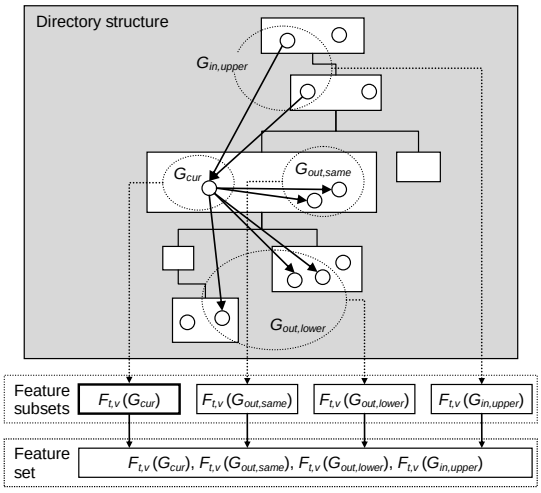

図５　周辺ページグループによる素性構成

を作成し，SVM による分類を行っている。周辺ページグループ構成としては，実験の結果から，対象ページ自体，及び，３種類の接続タイプと３種類のディレクトリ階層の個別の全組合せ（ただしディレクトリエントリと下位ディレクトリの組合せを除く）の，合計９個の周辺ページグループを用いることとした。

再現率（精度）保証型分類器の作成には，SVM のパラメータ（分類エラーコスト及び正例対負例の重み）を用いて再現率（精度）の要求を満たすようにチューニングを行った。

図６に ResJ-2 を用いて実験を行った結果として得た基本分類器の性能を示す。“baseline”は各 Web ページを単体として用いた結果，“A_tr”及び“U_t”は対象ページ分類器のみ（周辺ページグループ手法）による結果，“A_t+3-dir”はさらに周辺ページ分類器（ノイズページフィルタリング）を追加した手法による結果である。この結果から，周辺ページグループ手法により中～高再現

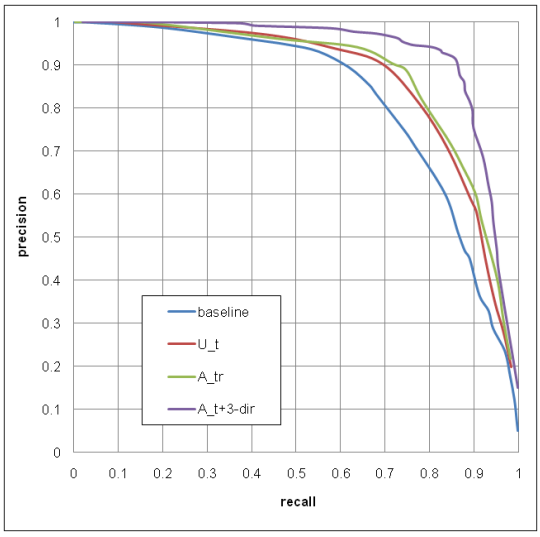

図６　基本分類器の性能

表１　NW100G-01 の３段階分類結果

		Requirement (precision/recall)		
		0.995/0.98	0.99/0.95	0.98/0.90
baseline	assured positive	4,776	8,097	12,490
	uncertain U_B	267,132	160,792	93,914
	assured negative	10,766,812	10,869,831	10,932,316
A_t+3-dir	assured positive	18,621	21,915	30,073
	uncertain U_A	165,849	75,939	27,913
	assured negative	10,854,249	10,940,866	10,980,733
Reduction ratio $\|U_A\|/\|U_B\|$		0.621	0.472	0.297

率（中～低精度）領域で，またノイズページフィルタリング手法により低～中再現率（高～中精度）領域で，それぞれ大幅な性能向上が得られることが示された。

本実験結果を適用して，３段階分類器により NW100G-01 を分類した場合の推定結果を表１に示す。未確定(uncertain)への分類結果は人手判定を要することになるが，要求される再現率及び精度のレベルに応じて，その Web ページ数は約 30～60%に減少させることができた。

WebKB を用いた基本分類器の実験結果においても同様に大幅な性能向上が確認されており，同じデータセットを用いた従来研究と比較しても十分に高い性能を実現している。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　４件）

① Yuxin Wang, Keizo Oyama: “Web page classification based on surrounding page model representing connection type and directory hierarchy”, 情報処理学会論文誌データベース, No.TOD-42, 2009, 印刷中（査読あり）

② Yuxin Wang, Keizo Oyama: “Building web page collections efficiently exploiting local surrounding pages”, Progress in Informatics, No.6, p.27-39, 2009（査読あり）

③ 相澤彰子, 高久雅生, 大山敬三: “大規模データベースを利用したリンケージシステムの提案と実装”, 日本データベース学会 Letters, Vol.6, No.4, p.17-20, 2008（査読あり）

④ Yuxin Wang, Keizo Oyama: “Combining page group structure and content for roughly filtering researchers’ homepages with high recall”, 情報処理学会論文誌データベース, Vol.47, No.SIG 8, p.11-23, 2006（査読あり）

〔学会発表〕（計　６件）

① Masao Takaku, Akiko Aizawa, Yasumasa Baba: “Name disambiguation of Japanese researchers: a case study with statistics research community”, Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis (IASC2008), 2008 年 12 月 5 日, Yokohama, Japan
② Yuxin Wang, Keizo Oyama: “Web page classification exploiting surrounding pages with noisy page filtering”, The 2008 International Conference on Data Mining (DMIN2008), 2008 年 7 月 14 日, Las Vegas, Nevada, USA
③ Atsuhiro Takasu, Kenro Aihara, Taizo Yamada: “A smoothing method for a statistical string similarity”, IEEE Intl. Conf. on Information Reuse and Integration (IRI2007), 2007 年 8 月 13 日, Las Vegas, USA
④ Yuxin Wang, Keizo Oyama: “Framework for building a high-quality web page collection considering page group structure”, Joint 9th Asia-Pacific Web Conference, APWeb 2007, and 8th International Conference, on Web-Age Information Management, WAIM 2007, 2007 年 7 月 16 日, HuangShan, China
⑤ Yuxin Wang, Keizo Oyama: “Web page classification considering page group structure for building a high-quality homepage collection”, Third International Conference on Web Information Systems and Technologies (WEBIST 2007), 2007 年 03 月 03 日, Barcelona, Spain
⑥ Yuxin Wang, Keizo Oyama: “Web page classification exploiting contents of surrounding pages for building a high-quality homepage collection”, 2006 年 11 月 27 日, Kyoto, Japan

６．研究組織

(1)研究代表者

大山　敬三 (OYAMA KEIZO)
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：９０１７７０２２

(2)研究分担者

なし

(3)連携研究者

高須　淳宏 (TAKASU ATSUHIRO)
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：９０２１６６４８

相澤　彰子 (AIZAWA AKIKO)
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：９０２１６６４８

高久　雅生 (TAKAKU MASAO)
物質・材料研究機構・科学情報室・主任エンジニア
研究者番号：００３９９２７１