2L-01

情報要求を満たさない文書の判別モデル構築と情報検索への活用

菅沼ひかり† 塩井隆円‡ 波多野賢治†
†同志社大学文化情報学部 ‡同志社大学大学院文化情報学研究科

1 はじめに

情報過多と言われる現代において，インターネット上の情報はユーザの必要性の有無や情報の真偽を問わず様々な形で存在している．そのため，これらの情報の中からユーザが必要とする情報を素早く，また正確に取得することは困難となっている．このような問題を解決するツールの一つに検索エンジンがある．検索エンジンでは適合率，再現率，F 値の向上が求められ，ユーザの情報要求を満たす情報の検索や不要な情報であるノイズを省くことなどが課題とされている．従来の研究のほとんどは，ユーザが入力した検索語にふさわしい Web ページ (以下，正解文書) をいかに検索，提示するかという観点から行われている．一方で，ユーザの情報要求に合致しない Web ページ (以下，不正解文書) は数多く存在するため，正解文書を導き出すよりも，比較的容易に発見することができる．

そこで本稿では，検索結果の不正解文書群特徴を定量的に分析し，得られた特徴を利用することで，あらかじめノイズとなる不正解文書を検索対象から除外するための正解/不正解文書判別モデルの構築を行う．そして，構築したモデルを情報検索に適用させることで検索精度向上を図る．

2 関連研究

不正解文書の一つとして Web スパム文書が挙げられるが，その検出方法は主に 2 種類に分類される [1]．一つ目は，コンテンツベースの Web スパム文書の検出方法である．サイト内の単語，URL などのコンテンツに着目し，Web スパム文書の特徴を取得することで，その特徴に合致するサイト内の Web 文書を Web スパム文書と判断する．Web スパム文書の特徴として，通常のサイトに比べて異常な URL が含まれる，コンテンツ変化が速い，ページレイアウトの複製が行われているなどが挙げられている．

二つ目は，リンクベースの Web スパム文書の検出方法である．リンクベースの Web スパム文書検出は，Web ページ間のホップ数，共引用，類似性などの位相関係に着目し，グラフ構造を利用する．そのグラフデータの構造を用いてリンク解析を行い，Web スパム文書のリンク特徴を得ることにより，それらの検出を行っている．

また，不正解文書を省くためにいくつかの検索エンジンでは，検索オプション機能を利用することで，ユーザ自ら検索対象としないキーワードを設定できるほか，様々な検索を行うことが可能である [1]．

3 提案手法

本稿では不正解文書を予め検索対象から省くため，スパム検出で利用されているコンテンツそのものではなく，正解/不正解文書群の定量的特徴に着目する．定量的特徴を把握した後，Support Vector Machine (SVM) を用いた未知文書に対する正解/不正解文書判別モデルの構築を行う．SVM は機械学習アルゴリズムの一つであり，多くの手法の中でも最も認識性能が優れている学習モデルの一つと考えられている [2].

判別モデル構築には，国立情報学研究所 NTCIR プロジェクト提供の NTCIR-5 WEB 文書データセット [3] を利用する．このデータセットは約 1 億ページの Web ページに対し約 1,200 題の検索課題が設定されており，各課題に対して各 Web ページに正解/不正解判別がなされている．このデータセットにおける正解/不正解文書を用い，判別モデル構築のための分析を行う．

各 Web ページの定量的特徴として，文字数 (ひらがな，漢字，片仮名)，リンク数 (順リンク，被リンク)，形態素数，HTML タグ数 (World Wide Web Consortium[2] で定義されているタグ)，マルチメディア数 (Apache HTTP Sever 2.4.16[3] で定義されている画像，動画，音声) を抽出した．ここで得られた定量的特徴に対して，t 検定を行うことで正解/不正解文書の定量的特徴における差を確認し，また，ロジスティック回帰分析の結果を用い，AIC によるモデル選択を行った．その結果，2 群間でいくつかの要素に有意差が見られた．例えば，出現する

Constructing A Discriminant Model for Unsuitable Retrieval Results and Its Application
Hikari SUGANUMA† Takamitsu SHIOI‡ Kenji HATANO†
† Faculty of Culture and Information Science，Doshisha University
‡ Graduate School of Culture and Information Science，Doshisha University

[1]Google 検索オプション http://www.google.co.jp/advanced_search (2015/08/24 閲覧)
[2]World Wide Web Consortium http://www.w3.org/ (2015/10/14 閲覧)
[3]Apache Software Foundation http://www.apache.org/ (2015/11/05 閲覧)

品詞の種類は全ての品詞において不正解文書の方が有意に多く，順リンクは不正解文書，逆リンクは正解文書が有意に多かった．この t 検定とロジスティック回帰分析，AIC によるモデル選択の結果から，以下の 3 パタンにおいて SVM による学習を行い，判別モデルの構築を行った．なお，AIC によるモデル選択は AIC 値に 2 以上の差があれば有意差があるとされるため，AIC 値が最小のモデルから差が 2 未満であるモデル四つを採用している．

1. 全ての定量的特徴を使用した場合 (model_all)
2. ｔ 検定において有意差がみられた定量的特徴を使用した場合 (model_t)
3. ロジスティック回帰分析及び AIC により選択された定量的特徴を使用した場合 (model_AIC1〜4)

4 評価実験

評価実験では各提案モデルの判別精度の確認を行った後，従来手法との比較，実際に情報検索に本提案モデルを適応した際の検索精度の確認を行う．

まず，モデル構築に未使用のデータから評価用データを作成し，各モデルを用いて判別を行った．判別結果と NTCIR-5 WEB で付与されている判別が一致する割合を判別精度とする．また，モデル構築に要する学習時間と判別に要する判別処理時間を測定した．その結果，model_AIC3 が最も有用であった．

従来手法との比較では，モデル構築に使用する定量的特徴の違いによる判別精度の差を確認する．従来のスパム検出では主に単語を文書の特徴とし，機械学習アルゴリズムの一つであるナイーブベイズを用い，判別を行っているが，モデル構築に使用する定量的特徴以外の条件を同じにするために，SVM を用いて単語の出現頻度に基づいたモデル (model_previous) を作成し，model_AIC3 と評価用データの判別を行った．

表 1 より model_previous と比較すると，model_AIC3 の方が全ての評価項目において有用である．このことから，正解/不正解文書の判別には単語以外の要素を考慮し文書判別を行う方が効果的であると判断できる．

次に，情報検索に提案モデルを適用した際の検索精度の確認を行う．汎用性を確認するためモデル構築に利用したデータではなく，NTCIR-3 WEB [4]，NTCIR-4 WEB [5] 文書データセットを利用し，Lemur Project 提供の indri[4] を用いて検索エンジンを構築した．同データセットの検索課題から無作為に課題を選択し，検索結果上位 100 件の適合率，再現率，F 値を算出した．

表 1: 従来手法との比較結果

モデル	判別精度 (%)	学習時間 (秒)	判別処理時間 (秒)
model_AIC3	74.70	0.790	0.0181
model_previous	59.37	631.59	79.2

表 2: 検索精度の評価

model_AIC3 適用の有無	適合率	再現率	F 値
model_AIC3 非適用	0.015	0.020	0.026
model_AIC3 適用	0.035	0.068	0.075

表 2 より提案モデル非適用の検索結果よりも，提案モデルを適用した検索結果の方が，適合率，再現率，F 値が向上した．このことから検索エンジンに本提案モデルを用いることは有用であると判断できる．しかし，3 値とも非常に低い値となっている．これは評価対象の文書数に対し，正解判定がなされている文書数が少ないことが原因と考えられる．

5 おわりに

本稿では，正解/不正解文書の定量的特徴を基に文書判別モデルの構築を行った．本提案モデルは従来手法よりも高い精度で文書判別が可能であり，情報検索においても検索精度が向上することが確認された．

今後の課題として，本提案モデルを組込んだ検索エンジンの構築を行う必要がある．

謝辞

本研究の一部は JSPS 科研費 15H02701 の助成を受けたものである．

参考文献

[1] Nikita Spirin and Jiawei Han. Survey on Web Spam Detection: Principles and Algorithms. *SIGKDD Explor. Newsl.*, Vol. 13, No. 2, pp. 50–64, 2012.

[2] Snehal S .Joshi and Navnath D .Kale. Survey : Support Vector Machine and Its Deviations in Classiction Techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 4, No. 12, pp. 993–998, 2014.

[3] Noriko Kando and Masao Takaku, editors. *Proceedings of the 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies : Information Retrieval, Question Answering and Cross-Lingual Information Access*. National Institute of Informatics, 2005.

[4] Oyama Keizo, Ishida Emi, and Kando Noriko, editors. *Proceedings of the 3rd NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering(September 2001-October 2002)*. National Institute of Informatics, 2003.

[5] Noriko Kando and Haruko Ishikawa, editors. *Proceedings of the 4th NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*. National Institute of Informatics, 2005.

[4] indri `http://www.lemurproject.org/indri/`