閲覧者の嗜好を考慮したWeb文書への広告挿入手法

鈴木督史† 鈴木優† 川越恭二†
†‡ 立命館大学 情報理工学部

1 はじめに

近年 Web においてコンテンツ連動型広告が普及している．コンテンツ連動型広告は，閲覧者の Web 文書に対する興味や関心に対して広告を挿入する．

広告挿入において，Web 文書に対する興味や関心が閲覧者間で異なるといった問題が考えられる．例えば，海外旅行に関する Blog 記事に対して広告を挿入する場合を考える．この場合，閲覧者が持つ興味や関心の対象は，イギリス，食事，観光地などのように閲覧者間で異なる．なぜならば，閲覧者間で嗜好が異なるためであると考えられる．このとき，既存の方法では，閲覧文書の内容による一様な広告を挿入するため，各閲覧者の興味や関心に適合する広告挿入は困難であると考えられる．

そこで本研究では閲覧者ごとに異なる嗜好を考慮した Web 文書への広告挿入手法を提案する．閲覧者の嗜好を考慮した広告を挿入することによって，閲覧者の Web 文書に対する興味や関心に適合する広告挿入が可能となる．

2 閲覧者の嗜好を考慮した広告手法

根本ら [1] は閲覧者の Web 閲覧履歴から閲覧者の興味の対象を抽出している．本研究において，この手法を用いて閲覧者の嗜好を抽出すると，複数の嗜好が同時に抽出される場合が考えられる．抽出された嗜好の組み合わせによっては，閲覧者の嗜好に一致しない場合が考えられる．例えば大阪と野球の嗜好が抽出されたからといって，必ずしも閲覧者が大阪の球団に対して興味を持っているとはいえない．また，嗜好の積集合が存在しない場合も考えられる．すなわち，閲覧者の嗜好を適切に反映した広告を挿入するためには，閲覧者の持つ複数の嗜好が適切に分類されることが必要である．

そこで本研究では，正しい組合せの嗜好を抽出するために，ソーシャルブックマーク（SBM）を利用する．SBM において閲覧者は，閲覧文書に対してタグと呼ばれるメタデータを付与してブックマークと分類を行い，これらの情報を Web 上で共有する．このようなタグ付けの特色から，SBM のブックマーク情報は，閲覧者の嗜好に適合する Web 文書を閲覧者の主観に基づいて分類した文書集合群であると言える．この文書集合群を利用することによって，閲覧者に適合した嗜好の組合せの抽出が可能となると考えられる．

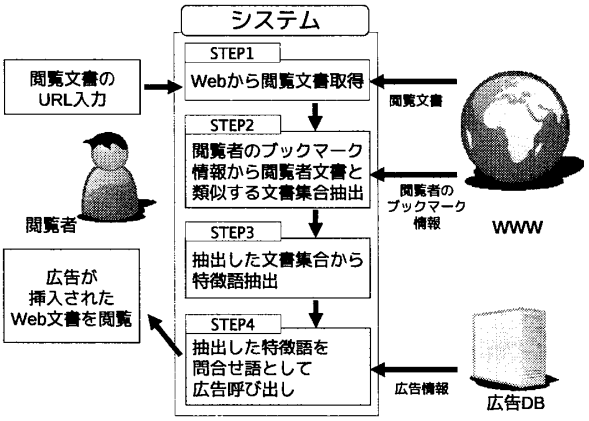

図 1: 広告挿入システムの概要

2.1 システム概要

本提案手法の有用性を検証するために広告挿入システムを開発した．システムが閲覧文書に対して広告挿入を行う手順について図 1 に示す．閲覧者はシステムに対して閲覧文書の URL を入力する．URL が入力されるとシステムは **STEP1** として Web から閲覧文書を取得する．次に **STEP2** においてシステムは閲覧者のブックマーク情報から，最も閲覧文書と類似度の高い文書集合を抽出する．この文書集合を本研究では特徴文書集合と呼ぶ．特徴文書集合は，閲覧文書と類似する閲覧者の嗜好の特徴を表すと考えられる．**STEP3** では特徴文書集合から特徴語を抽出する．最後に **STEP4** として **STEP3** において抽出した特徴語を問合せ語として広告を呼び出し，その広告を閲覧文書に対して挿入し，閲覧者に表示する．

An Advertisement Insertion Method based on User's Preference

Masafumi SUZUKI†, Yu SUZUKI‡and Kyoji KAWAGOE‡
†‡College of Information Science and Engineering,Ritsumeikan University
†msuzuki@coms.ics.ritsumei.ac.jp
‡{yusuzuki,kawagoe}@is.ritsumei.ac.jp

2.2 特徴文書集合抽出

システムは類似度算出のために，Web 文書に対して特徴ベクトルを付与する．本研究では Web 文書に対する特徴ベクトルとして単語の出現頻度を用いる．Web 文書 i の特徴ベクトルを (1) 式に示す．

$$\boldsymbol{P}_i = [f_i(w_1), f_i(w_2), \cdots, f_i(w_t)] \ (0 \leq i \leq n) \quad (1)$$

$\boldsymbol{P}_i$ は Web 文書 i のベクトル，n はシステムが扱う Web 文書数を表す．また，$f_i(w_t)$ は，Web 文書 i での単語 w_t の出現頻度，t はシステムが扱う単語数を表す．システムは，Web 文書 i からタグやタイトルなどの本文以外の情報を取り除き，その結果に対して形態素解析することによって単語を抽出する．

次にシステムは，閲覧文書と閲覧者のタグ付けによって分類された Web 文書の文書集合間のベクトルを用いることによって類似度を算出する．文書集合の特徴ベクトルは，文書集合に含まれる Web 文書の特徴ベクトルの総和とする．比較する方法として TF-IDF 法 [2] による各要素の重みを用いる．ここで求めた類似度の最も高い文書集合を特徴文書集合とする．本研究では，類似度の定義としてコサイン尺度を用いる．

2.3 特徴語抽出

2.2 節において抽出した特徴文書集合から特徴語を抽出する．本研究では特徴語を抽出するために，TF-IDF 法を用いることによって，文書集合に含まれる各単語に対して重みを与える．重みの値が大きいほど，その単語は特徴文書集合の特徴を表していると判定することができる．単語 w_t の重み $S(w_t)$ は，以下の式で表される．

$$S(w_t) = \sum_{i=0}^{m} f_i(w_t) \frac{N_{global}}{D_{global}(w_t)} \quad (2)$$

$f_i(w_t)$ は 特徴文書集合に含まれる文書 i に含まれる単語 w_t の出現頻度を表し，m は特徴文書集合の文書数を表す．また，$D_{global}(w_t)$ は，閲覧者のブックマークの文書集合における単語 w_t の文書頻度，N_{global} は閲覧者のブックマーク集合に含まれる文書数を表す．

特徴文書集合での IDF 値を用いると，この文書集合は閲覧者の主観による分類に基づいているため，特徴的な単語に対しても低い重みが付けられると考えられる．このことから，閲覧者のブックマーク情報の文書集合における IDF 値を利用した．

3 評価実験

本章では，提案システムの有用性を検証するための評価実験の方法について述べる．広告呼出しシステムとして Amazon Web Services^{TM1} を利用した．

3.1 実験方法

各閲覧者は興味のある 10 数個の Web 文書を事前に定義する．システムはこれらの Web 文書に対して，閲覧者の嗜好に基づいた広告の挿入を行う．本実験では，広告挿入手法として，3 種類の手法を用いた．

手法 A: 閲覧者のブックマークの分類を利用して嗜好を抽出

手法 B: 閲覧者のブックマーク全体を利用して嗜好を抽出

手法 C: 閲覧者の嗜好を利用しない

手法 A は，本提案システムである．手法 B は，閲覧者の分類を考慮せずに閲覧者のブックマークに含まれる文書全体から閲覧者の嗜好を抽出する手法である．手法 C は，閲覧文書において TF-IDF 値の大きい単語を，本提案システムにおける特徴語とする手法である．この際の IDF 値はシステムが扱うすべてを対象として求める．

広告挿入結果について，以下に示す主観評価を行う．

ケース 1 : 閲覧文書の内容で閲覧者の閲覧文書に対する興味関心に適合

ケース 2 : 閲覧文書の内容だけに適合

ケース 3 : 閲覧者の嗜好に適合するが，閲覧文書の内容に適合せず

ケース 4 : いずれにも適合せず

4 おわりに

本研究では，閲覧者ごとに異なる嗜好を考慮した Web 文書への広告挿入手法を提案した．今後は 3.1 節の実験を行い，実験結果について考察することによって，本提案手法の有用性を検証する．

参考文献

[1] 根本潤, 遠山元道：“閲覧履歴に基づく情報検索の相互支援”, 電子情報通信学会, DEWS2004, 3-B-02 (2004).

[2] G. Salton Ed.: “Automatic Text Processing”, Addison-Wesley Longman Publishing Co., Inc. (1985).

[1] http://developer.amazonwebservices.com/