D-011

タグ組み合わせに基づく Web コンテンツ検索
Search of web contents based on combined tags

福盛秀雄† 村岡洋一†
Hideo Fukumori Yoichi Muraoka

1. はじめに

Web 上システムにおいては近年、コンテンツの提供者あるいは閲覧者がタグと呼ばれるキーワードを付加し、これをコンテンツのナビゲーションに利用する手法が広く用いられている。閲覧者がこれらタグを元に情報を拾い出そうとする際には、例えばタグクラウドと呼ばれるタグの一覧を提示したり、あるいは特定のコンテンツに付けられているタグの集合を提示し、これをハイパーリンクなどを経由して検索するかたちでの情報取得方式が主流である。

これらはいずれも単一のタグに基づいた取得方式であるが、結果として表示コンテンツの数が膨大となり、その把握が困難となること、また、単一のタグはあいまいとなる傾向があり、ユーザが望むコンテンツを示すには不十分であるという問題がある。

本発表ではこれらの問題に対し、複数のタグの組み合わせによるコンテンツ検索の方式を提案する。既存の研究としては共起度に基づき「タグクラスタ」と呼ばれる集合を提示する方式などが提案されているが、今回提案する方式は件数絞り込みの側面に着目し、効果的な絞り込みを行うことを目的とする。

2. 既存のタグシステムと問題点

2.1 既存のタグシステム

2.1.1 ナビゲーションにかかわる問題点

現在のタグシステムの問題点として一つ挙げられるのは、一つのタグを選択した際に示されるコンテンツの候補件数が数万から場合によっては数百万のオーダーとなることである。例えば写真アップロードサイトとして有名な flicker[1]において、”wedding”というタグの付いた写真を選択すると、約 1000 万件の写真が表示される。

このような大量のコンテンツを表示する際にはページングなどの方式が一般的にとられるが、例に挙げたような多数のコンテンツに対するナビゲーションとしては実際に機能するものとは言い難い状況にある。

2.1.2 絞り込みによる効果

一方、flicker や del.icio.us[2]に代表される既存のソーシャルブックマークシステムにおいては、タグを複数 AND 条件で結ぶことにより絞り込みを行う機能は提供されている。

複数タグを用いた絞り込みは、多量のコンテンツに対するナビゲーション手段として有効な対策となり得る。例えば先の例で挙げた flicker 上での”wedding”タグの付けられた写真は約 1000 万件であるが、”wedding”と”church”タグの両者が現れる写真は約 11 万件となる。さらに”wedding”, “church”, “stanford”の三つのタグが付けられた写真で絞り込みを行うと 280 件までの絞り込みが行われる。ちなみに”church”単体のタグの付けられた写真は約 220 万件、”stanford”単体のタグの付けられた写真は約 12 万件存在している。この例は、個別のタグ自体で十分な絞り込みができないものが、複数のタグを組み合わせることによってナビゲーションが容易となるレベルまでの件数に絞り込まれることを示している。

† 早稲田大学 理工学術院 Faculty of Science and Engineering, Waseda University

2.2 タグシステムの従来研究と問題点

共起グラフを元に関連の高いタグの集合を取り出したものをタグクラスタと呼び、これをナビゲーションの助けとする方式については、従来より多く研究されている。Begelman, et al[3]はあるタグと強く関連するタグの集合を提示するための方法として、タグの共起度を元にクラスタを自動的に構成し、これを文脈（セマンティック）レベルで切り離すことによりタグの集合を提示する方式を提案している。

一方で、絞り込みという観点からこういったタグクラスタに基づく方式を見た場合には不十分に見える点も多い。具体的には

- 共起性の高いタグの集合を集めることを中心に据えているため、絞り込みに有効となるタグの組み合わせは共起性が低いものとして除外されることが多い
- コンテンツの集合の件数を示すことができない

といった点が挙げられる。

一方で、情報推薦の一手法である collaborating filtering を用いて、あるユーザと嗜好の一致するコンテンツ群を提示するアプローチも多く提案されている[4][5][6]が、やはり絞り込み型ナビゲーションの手段としては同様に不十分である。

3. 提案手法

タグシステムにおいて、特定のタグを選択した場合に、そのタグにより提示されるコンテンツ群に含まれるタグの一覧を提示し、かつ選択されたタグと提示されたタグを AND 条件で絞り込んだ件数を提示することにより、効果的なナビゲーションを行う。

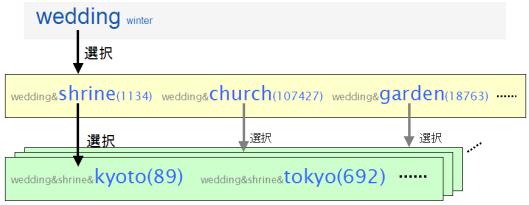

図 1 絞り込み候補 提示インタフェース

提示インタフェースにて、AND 条件で結ばれたタグの件数を表示するためには事前に計算処理を行う必要があるものと考えられる。事前計算の方法としては、

1. あるタグを持つコンテンツの集合を、重なりを持たない共起タグの集合に分割し、同時にその件数を算出する。（図 2）
2. 分割された部分集合のそれぞれについて、上記 1.の操作を繰り返し行う。（図 3）

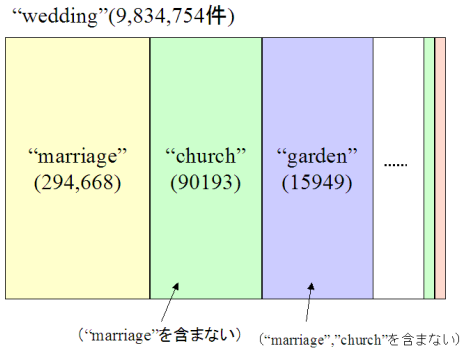

図 2 提示タグの算出（１）

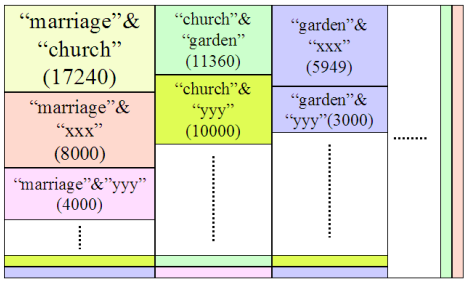

図 3 提示タグの算出（２）

4. 複数タグ絞り込み方式の妥当性に関する検討

実装に先立ち、論文ソーシャルブックマークサイトである citeulike[7]を対象に、複数タグの組み合わせの妥当性について検討した。

対象として、論文ブックマークサイト citeulike のタグ情報を元に、”software”タグの付けられたコンテンツ（総件数 7650 件）に対し、さらにタグでこれをさらに絞り込む上で提示されうるタグの数について確認を行った。

その結果、提示されうるタグ数は 663 個、絞り込みで示されるコンテンツ件数の最大値は”software”&”engineering”の組み合わせによる 1010 件、コンテンツ件数が 1 件のみとなる組み合わせは 342 個という結果を得た。あくまで予備的な調査であるが、提示タグ候補の過半数において、コンテンツ件数が 1 件のみとなる組み合わせとなる点などについてはさらに対応を検討する必要があるものと考えられる。また、タグの検証範囲の拡大、前述の flickr や del.icio.us など他のシステムでの検証も行う予定である。

5. まとめ

大量のコンテンツを含むシステム上において、絞り込みを効率的に行い目的とするコンテンツを取得する方法として、複数のタグの組み合わせによる提示の方式を示した。

今後は、

- タグ絞り込みの効果についてのさらなる検証
- 絞り込み件数の提示のための効率的な計算方法の検討
- 実装による検証

を進めていく予定である。

参考文献

[1]flickr, http://flickr.com
[2]del.icio.us, http://del.icio.us
[3]G. Begelman, P. Keller, and F. Smadja, "Automated tag clustering: Improving search and exploration in the tag space," Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, (2006)
[4]百田 信, 伊藤 栄典, “ソーシャルブックマークに基づく情報発見”, DEWS2008, (2008).
[5]佐々木祥, 宮田高道, 稲積泰宏, 小林亜樹, 酒井善則, "Social Bookmark におけるコンテンツクラスタ間の類似度を用いた web コンテンツ推薦システム," 情報処理学会論文誌:データベース, Vol.48, No.SIG20(TOD36), (2007)
[6]佐々木祥, 宮田高道, 稲積泰宏, 小林亜樹, 酒井善則, "Folksonomy におけるコンテンツ推薦のためのメタデータ成長モデルの提案, 情報処理学会研究報告. データベース・システム研究会報告, Vol.106, No.150, (2006)
[7]citeulike, http://www.citeulike.org