2006－NL－171（3）
2006／1／12

個人文書アーカイブを用いた文書有用性判定タスクとそのための類似判定

漆畑 航太　　梅村 恭司

豊橋技術科学大学 情報工学系

WWW におけるブラウジング支援手法のひとつに、個人化 (Personalizing) がある。個人化とはユーザーの持つ文書や行動の情報から、よりユーザーにとって有用そうな、ユーザーが興味を持ちそうな情報へ辿り着きやすくする技術である。ブラウジングの個人化にあたっては、未知の Web ページがどの程度有用であるかを判定する必要がある。本研究ではそのためのタスク、「文書有用性判定タスク」を規定し、それに対する文書ベクトル空間を利用したアプローチを示す。さらに、この文書有用性判定タスクのアプローチを実装したブラウジング支援システムについて示す。

Usefulness judgment task with individual document archive and similarity judgment for that task.

Kota URUSHIBATA　　Kyoji UMEMURA

Toyohashi University of Technology (Information & Computer Sciences)

Personalizing is one of the assistance scheme for WWW browsing. In personalizing, system will hopefully provide usefulness information for desirable documents. Since the system need to guide users to more useful or interest page for users. To make it possible, the system need to judge that a desirable document is useful or not. In this research, we define a “usefulness judgment task” and show the approach using document vector space. We also describe a browsing assistance system that implements our approach.

1　はじめに

World Wide Web(WWW) は近年では広く普及し、一般的な情報源となっている。巨大な WWW から必要な情報を探すために、Yahoo! のようなポータルサイトや、Google のような検索エンジンが用いられている。

一方で情報探索を容易にするためのアプローチとして、ブラウジングの個人化 (Personalizing) がある。ユーザの持つ情報や過去の行動をもとに、個人個人に応じた Web ページの表示などのブラウジング支援を試みる技術である [1],[2]。

ブラウジングの個人化にあたっては、新しい (これから見ようとする)Web ページがどの程度有用であるかを判定する必要がある。本研究ではそのための「有用性判定タスク」を考える。有用性判定タスクを規定し、それに対するアプローチについて示す。また、実際に構築したシステムについて実行例を示す。

2　文書有用性判定タスク

まず、ここでは文書有用性判定タスクがどのようなものであるかを述べる。

文書有用性判定タスクでは 2 種類の文書を用いる。まずひとつは有用性判定の根拠となる「個人文書アーカイブ」である。個人文書アーカイブは特定の話題を扱った文書の集合である。もうひとつは有用性判定の対象となる「対象文書」である。

文書有用性判定タスクは個人文書アーカイブの内容に基づいてある対象文書が有用であるかどうかについて Yes 又は No の判定を下す。

ユーザが持つ個人文書アーカイブを「ユーザが興味を持っている分野の文書集合」、有用性を「ユーザが興味のある分野の情報でありそうかどうか」と言い換えると、このタスクは個人文書アーカイブの取り扱う特定の話題群との類似判定のタスクと考えることもできる。

2.1　他の類似判定タスクとの比較

文書の類似性を判定するタスクに、情報検索での類似判定やスパムフィルタリングなどがある。本節では有用性判定のタスクとこれらとの違いを説明する。

これらのタスクの違いは、取り出したい正解の範囲の違いにある。。

2.1.1　情報検索の類似判定タスク

情報検索の類似判定ではクエリとなる文書に類似する文書を検索したい文書、即ち正解と判定する。

情報検索は検索対象の全体からクエリに相応しい文書を探すタスクである。そのため、正解となる文書はクエリに近い比較的狭い範囲にある文書の集合となる。全体の文書の範囲と正解文書の範囲のイメージを図 1 に示す。

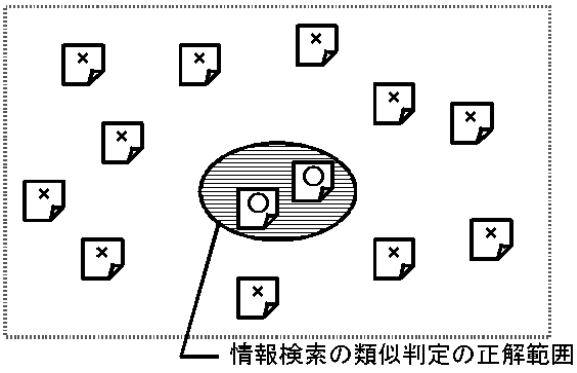

図 1　情報検索の類似判定での正解範囲

2.1.2　スパムフィルタリング

スパムフィルタリングはメールの集合からスパムメール (ユーザにとって必要のないメール) を取り除くタスクである [3]。まず、判定の根拠となる学習用データとしてスパムメールとスパムではないメールの集合を与える。そして、学習したデータをもとにスパムメールかどうかを判定する。

スパムフィルタリングでは文書全体から狭い範囲にあるスパム文書を取り除いた残りの範囲が正解となる。そのため、正解は比較的広い範囲に偏在する。全体の文書の範囲と正解文書の範囲のイメージを図 2 に示す。

2.1.3　文書有用性判定タスクとの正解範囲の違い

今回取り扱う文書有用性判定タスクでは、個人文書アーカイブが取り扱う特定の話題に含まれる文書を取り出す。正解となる文書は個人文書アーカイブの話題に近い範囲となる。個人文書アーカイブは多数存在するため、正解となる範囲は特定の類似文書だけ取り出す情報検

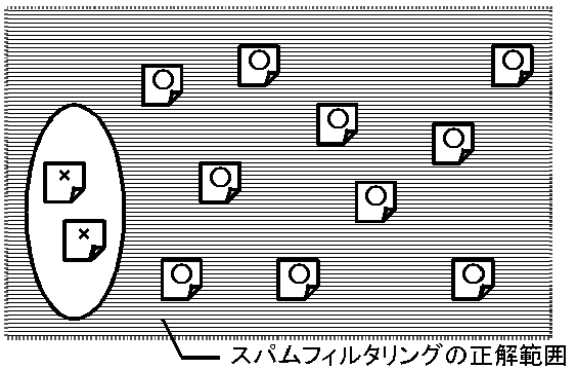

図 2　スパムフィルタリングでの正解範囲

索より広く、全体から特定範囲だけ取り除いたスパムフィルタリングよりは狭くなると考えられる。全体の文書の範囲と正解文書の範囲のイメージを図 3 に示す。

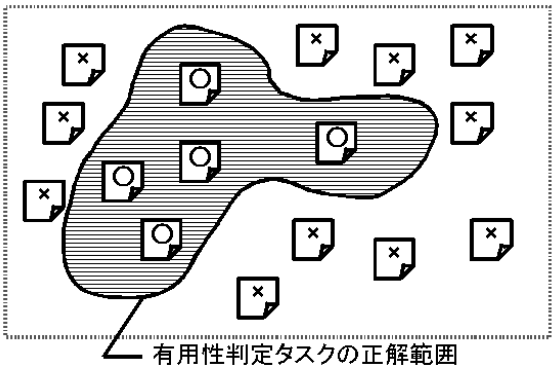

図 3　文書有用性判定タスクの正解範囲

3　アプローチ

ここでは文書有用性判定タスクへのアプローチについて述べる。本研究では文書有用性判定のために、文書ベクトル空間を利用した類似性判定を行う。

3.1　文書ベクトル空間

今回のアプローチでは文書ベクトル空間を利用する。個人文書アーカイブに登場する全ての単語の種類数を n とする。このとき、ある文書のベクトルは $\overrightarrow{d} = \{w_1, w_2, \cdots, w_n\}$ と表すことができる。それぞれの w_i は文書中での出現回数である。

3.1.1　文書のベクトル化

文書をベクトル化するにあたっては、単語の切り出しが必要となる。英語の文書は単語が空白で区切られているために切り出しは比較的容易だが、日本語の文書は単語の切れ目を空白で判断することができない。

日本語の文書から単語を抽出する方法は、形態素解析や Bigram などが考えられている。今回はメールや Web ページを取り扱う文書として想定するため、文書構造の崩れた文書が出現する可能性がある。そこで意味的構造に頼らない方法で単語を取り出す方法を選択した。

“Term” と呼ぶ短い部分文字列を取り出し、文書ベクトル空間で用いる単語として取り扱う。本研究では「ひとつながりの英字」「ひとつながりのカタカナ」を Term とし、和文からは漢字の Bigram を Term として取り出した (図 4)。

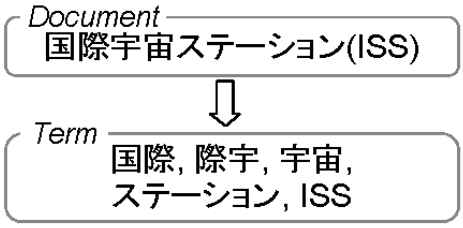

図 4　Term の切り出し例

3.1.2 距離の概念

ここでは、文書ベクトル同士の距離について述べる。文書ベクトル空間での距離の概念にはいくつかあるが、今回はユークリッド距離を使用した。

$$distance(d_1, d_2) = \sqrt{(w_{11} - w_{12})^2 + \cdots + (w_{n1} - w_{n2})^2} \quad (1)$$

ここで w_{ij} は文書 d_j での単語 w_i の出現回数を表す。

3.2 $threshold$ の決定

まず先に個人文書アーカイブを使用して $threshold$ を決定する

1. 個人文書アーカイブ $D = \{d_1, d_2, \cdots, d_n\}$ に含まれる文書全てについて、自分以外の最も距離の近い文書 (最近傍:Nearest Neighbor) を計算する。文書 d_i から最近傍までの距離を $dist_i$ とする。
2. ここで、それぞれの文書の最近傍までの距離の集合を $Dist = \{dist_1, dist_2, \cdots, dist_m\}$ とする。
3. この集合 $Dist$ を昇順にソートし、距離集合の数列を得る。
4. $Dist$ の $\lfloor 0.95m \rfloor$ 番目の要素値を $threshold$ にする。

3.3 有用性判定

有用性判定では、判定対象の文書に Yes または No の判断をする。

1. 個人文書アーカイブ D に含まれる文書ひとつひとつについて、対象文書 q との距離 $dist(q, d_i)$ を計算する。
2. 最も小さい $dist(q, d_i)$ の値 (即ち、対象文書から個人文書アーカイブへの最近傍) が $threshold$ より小さければ Yes、大きければ No と判定する。

3.4 アプローチの特徴

このアプローチでは、個人文書アーカイブ全体の Term の出現情報だけを参照するという方法ではなく、対象文書と個人文書アーカイブに含まれる個別の文書との最近傍を計算する。

これは、個人文書アーカイブが多数の文書からなり、有用性判定タスクの正解の範囲はそれらの文書の近くに広がっているという考え方に基づいている。

4 システムとしての実装例

本研究では、文書有用性判定タスクを検討することに加えてブラウジング支援を行うためのシステムを開発した。

システムでは、個人文書アーカイブの実体はユーザの興味ある話題のメールの集合となる。また、対象文書は表示する Web ページに含まれるハイパーリンク先の Web ページとなる。

システムは大きく分けてふたつのプログラムから成る。ひとつは個人文書アーカイブから $threshold$ を決定するプログラムであり、もうひとつは実際にブラウジング支援を行うプロキシプログラムである (図 5)。

4.1 $threshold$ の決定

個人文書アーカイブから $threshold$ 決定する部分では、前述 (Section3) のアプローチで説明した $threshold$ 決定アルゴリズムに相当する

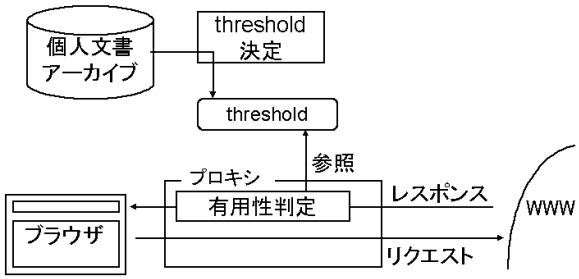

図 5　システム模式図

動作をする。

4.2　プロキシプログラム

実際にブラウジング支援を行うプロキシプログラムは次のような手順で有用性判定タスクを行う。

1. ブラウザからリクエストを受け取る
2. リクエストの URL にアクセスし、ページを取得する
3. ページに含まれるハイパーリンクを抽出する
4. 抽出したハイパーリンクの指し示すページの文書を、対象文書として Section3 で示した有用性判定を行う
5. 有用性判定の結果情報ををはじめにリクエストされたページ内に埋め込み、ブラウザへのレスポンスとする

つまり、システムは表示する Web ページ内にあるハイパーリンク先を先読みする。これによって、今見ているページのハイパーリンク先についての有用性の情報をユーザから開くことなく確認することができるようになる。

4.3　システム動作例

システムの動作を確認した動作例を示す。

4.3.1　動作実験環境

個人文書アーカイブとして、VineLinux のメーリングリスト[*1]のメール 1000 通を用いた。

有用性判定タスクの結果は、表示した Web ページに含まれるハイパーリンクの横に付与する。結果が Yes であれば「○」を、No であれば「×」を挿入する。

4.3.2　動作実験結果

図 6 はシステムを適用する前の Web ページの表示結果であり、図 7 はシステムを適用した場合の表示結果である。

システムを適用した結果、Web ページ内のハイパーリンク横に有用性判定タスクの結果が「○」と「×」で付加されていることが分かる。

4.3.3　動作実験の考察

実験によって得られた有用性判定タスクの結果では、目視での Web ページと個人文書アーカイブとの類似性と判定結果と相関性は見られなかった。

文書の類似度を測る際の距離の概念や、Web ページからの情報の取り出し方法などのアプローチについて改善する必要があると考えられる。

5　まとめ

今回はブラウジング支援のための文書有用性判定タスクについて規定し、そのためのアプローチを示した。また、そのアプローチを組み込んだシステムの動作実験について示した。

今後は前述の通りアプローチの見直しや、システムの改善、また、有用性判定タスクの評

[*1] http://vinelinux.org/ml.html

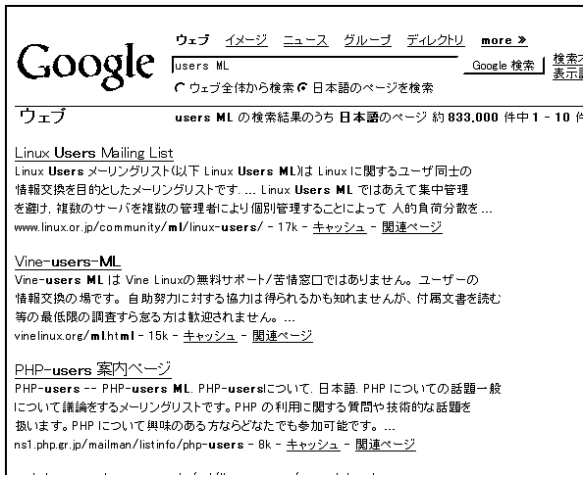

図 6　動作例:適用前

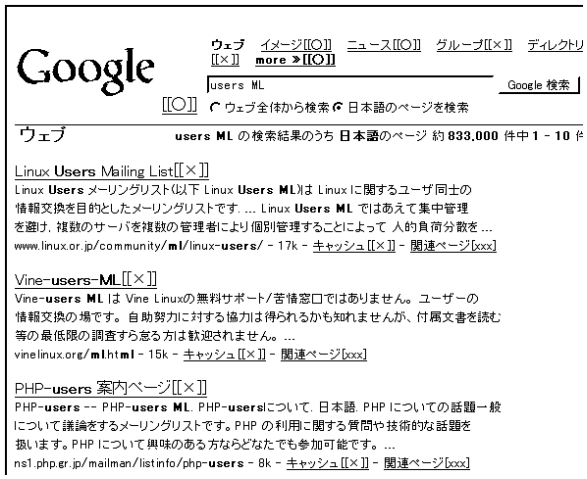

図 7　動作例:適用後

価をする方法を確立することが重要と考えられる。

参考文献

[1] B. Mobasher, R. Cooley and J. Srivastava. Automatic Personalization Based On Web Usage Mining *Communication of ACM, Volume 43 Issue 8*, 8 2000.

[2] 松尾 豊、福田 隼人、石塚　満., ユーザ個人の閲覧履歴からのキーワード抽出によるブラウジング支援 **人工知能学会誌**, *Vol.18, No.4E, pp.203-211* 2003.

[3] Paul Graham. A Plan for Spam. *http://www.paulgraham.com/spam.html*, 8 2002.

[4] 北 研二, 津田 和彦, 獅々堀 正幹. 情報検索アルゴリズム. **共立出版**, 1 2002.