D_048

メールテキストを用いた検索エンジンの検索結果の並び替え

An Investigation into Personalizing Web Search Results using Mail Texts

上田　洋† 福居　和男‡ 宮原　良一+‡ 村上　晴美+
Hiroshi Ueda Kazuo Fukui Ryoichi Miyahara Harumi Murakami

1．はじめに

インターネット上の情報が増え続ける中、検索エンジンの検索結果も膨大となっており、個人にとって必要な情報を迅速に得ることが難しくなってきている。これまで、Google のパーソナライズド検索など, Web 検索履歴を用いて検索エンジンの検索結果を並び替える研究が実用化されつつあるが、Web 検索履歴はユーザの雑多な情報要求に基づくものであり、特定のテーマの切り出しが難しい。そこで、本研究では、特定のテーマをゆるやかに切り出すことのできる情報源として、メールテキストの利用を検討する。

本研究では、ユーザが持つメールテキストからユーザプロファイルを作成し、検索エンジンの検索結果を並び替える手法を検討する。

2. 提案手法

提案手法は、メールログからのユーザプロファイルの作成、検索結果の並び替え、の2段階から構成される（図1）。

まず、ユーザの指定したメールログからメールテキストを抽出してユーザプロファイルを作成しておく。

ユーザが文字列を入力すると検索エンジンで検索を行い、Web ページを 100 件取得する。ユーザプロファイルと Web ページ間の類似度を余弦を用いて計算し、類似度の高い順に並べ替えて出力する。

図2にシステムの画面例を示す。画面上部に抽出したメールテキスト一覧、画面下部に検索質問「人工知能」に対する Google による検索結果と、本手法による並び替え結果を示す。なお、この例は、実験（後述）におけるメールログを利用して検索結果を並び替えたものである。

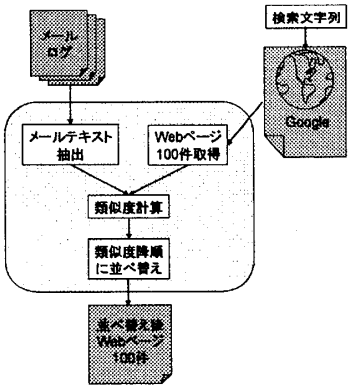

図1：手法概要

図2：画面例

2.1 ユーザプロファイルの作成

(1) メールログの前処理

メールログはメールクライアント（以下、メーラ）によって保存形式が異なる。本研究では、多様なメーラに対応するため、メールサーバに保存される形式としてよく用いられる2種類のログ形式に対応することとした。一つは、複数のメールを一つのファイルとして保存する mbox 形式である。もう一つは、一つのメールを一つのファイルで保存する Maildir 形式である。本研究では、mbox 形式のメールログを Maildir 形式に変換しておく。

(2) メールテキストの抽出

Maildir 形式のメールログから、件名（Subject 行の内容）、本文、To 行に記述されたメールアドレス（以下、To アドレス）、From 行に記述されたメールアドレス（以下、From アドレス）を抽出してメールテキストとする。

件名、To アドレス、From アドレスについては、ヘッダから抽出を行う。件名に関しては、日本語文字列の抽出に対応するため、MIME エンコードのデコード処理を抽出と同時に行う。本文については、シングルパートのメールログについては、本文全てを抽出する。マルチパートのメールログについては、BASE64 でエンコードされていないテキストデータのみを抽出し、BASE64 でエンコードされたバイナリデータ、テキストデータを除去する。

なお、本手法では、メールテキスト一覧を画面表示することにより、ユーザがユーザプロファイル作成に用いるメールテキストを選択する機能を備えている。

†大阪市立大学大学院工学研究科
‡インターネット株式会社
+大阪市立大学大学院創造都市研究科

(3) メールベクトルの作成

メールテキスト中の件名、本文から、形態素解析システム茶筌を用い、2 文字以上から構成される名詞を索引語として抽出し、その頻度を重みとするメールベクトルを作成する。

2.2 検索結果の並び替え

ユーザが文字列を入力すると、Google WEB APIs を利用して、100 件の Web ページのタイトル・URL を取得し、各 URL から、Web ページのテキストデータを取得する。このテキストデータから、メールテキストと同様に、形態素解析により 2 文字以上から構成される名詞を抽出して頻度計算を行い、Web ベクトルを作成する。

Web ページから作成した Web ベクトル、メールテキストから作成したメールベクトル、のベクトル間の余弦を用いて、類似度を計算する。類似度の高い順に Web ページ 100 件を降順に並び替える。

ベクトルの余弦の計算式は、以下の通りである。

$$sim(d_x, d_y) = \frac{\sum_{i=1}^{T} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{T} x_i^2 \times \sum_{i=1}^{T} y_i^2}}$$

なお，索引語は d_x, d_y に出現する 2 文字以上から構成される名詞であり、重みはその頻度である。

3. 実験

3.1 方法

大阪市立大学大学院創造都市研究科の情報学系研究室のメーリングリスト（以下、ML）を用いた。ML は、2005 年 7 月 15 日に開始され、メンバーは、2006 年 6 月 21 日現在、教員 1 名、大学院生 8 名の計 9 名である。実験には 2005 年 7 月 15 日から 2006 年 5 月 11 日の間にやり取りされたメール計 134 通を用いた。被験者は、ML メンバーの大学院生 5 名である。

検索質問は「インターネット」「インタフェース」「情報検索」「人工知能」「図書館」「大阪市立大学」の 6 件とした。最初の 5 件は研究室の研究テーマに関連する。

各検索質問に対して、Google の結果 10 件と本手法の並べ替えで得られた Web ページ上位 10 件を取得した。その後、元の順位や、どちらから得た結果かをわからないようにするために、結果をマージしてランダムに並び替え、最大 20 件のリストを作成した。リストの各タイトルを選択すると Web ページを見ることができる。

被験者には、リストのタイトルを選択して Web ページを閲覧後に評価をさせた。

評価尺度として、「関連度」、「有用性」、「新規性」を用いた。「関連度」は、ML と各 Web ページがどの程度関連しているかを 3 段階（3.非常に関連している、2.関連している、1.関連していない）で評価させた。「有用性」は、被験者が各 Web ページをどの程度有用であると考えるかを 3 段階（3.非常に有用である、2.有用である、1.有用でない）で評価させた。「新規性」は、被験者が今まで見たことがある Web ページを 0, 見たことがないページを 1 で評価させた。

3.2 結果と考察

関連度と新規性については、本手法の方が結果が良かった（関連度、本手法：1.87、Google：1.77、新規性、本手法：7.00、Google：5.70）が、有用性については、Google の方が良い結果が得られた（本手法：1.59、Google：1.68）（表 1 参照）。

関連度と新規性について本手法と Google との間に有意差が見られた（関連度：$t(29) = 2.60$（$p<.05$）、新規性：$t(29) = 4.86$（$p<.01$））。

本手法と Google の出力された Web ページのうち、両者とも同じものの数は、各検索質問につき 1 ないし 2 ページと少なかった。

関連度については、本手法がメールテキストの内容を並べ替えに反映できた結果であると考える。

新規性については、本手法を用いた方が、被験者が見たことのないページを多く得ることができると考える。

有用性については、Google の方が数値は高い。しかし、本手法と Google の間には有意差がなく、数値の差も少ないため、有用性についても一定の水準のページが出力された結果であると考える。

表 1：関連度、有用性、新規性

	関連度*	有用性	新規性**
Google	1.77	1.68	5.70
提案手法	1.87	1.59	7.00

*: $p<.05$, **: $p<.01$

4. 関連研究

メールテキストを用いて Web 情報を提示する研究が存在する（たとえば[1,2]）。文献[1]では、ユーザの置かれた状況では思いも付かない意外性のある Web ページを推薦する。文献 [2] では、メールテキストの内容を基に関連情報を推測し Web ページの検索を行い、Web 上に存在する関連情報を提示する。本研究は、Web ページを提示する点において上記と類似するが、検索結果を並び替える点が異なる。

5. おわりに

本研究では、ユーザのメールテキストを用いて検索エンジンの検索結果を並び替える手法を検討した。情報系研究室のメーリングリストを用いた評価実験の結果、メールテキストの内容を反映した検索結果を得ること、ユーザが過去に見たことがないページを見ることができることが示唆された。今後は、ユーザプロファイル作成手法の改良を試みるとともに、多様な文脈におけるメールテキストや検索質問を用いた評価実験を行いたいと考えている。

参考文献

[1] 斎藤真理, 山本則行, 電子メールからの興味抽出手法と意外な Web ページと出会うきっかけを与えるエージェントシステム, 認知科学, Vol.11(3), pp.252-261, 2004.

[2] 赤星祐平, 小山聡, 角谷和俊, 田中克己, 携帯端末による電子メール交換に基づく Web 検索, 日本データベース学会 Letters, Vol.2, No.1, pp.111-114, 2003.