Web検索エンジンに対する検索語の類似度に基づく関連文書の検索

D-8 Related Document Retrieval based on Similarity between Web Search Words

安川 美智子† 山田 篤†‡ 星野 寛†‡ 大瀬戸 豪志†† 上林 彌彦†
Michiko Yasukawa Atsushi Yamada Hiroshi Hoshino Takashi Oseto Yahiko Kambayashi

1．まえがき

インターネット上の膨大な情報の中から，少数の有用な情報を探し出すことは難しく，時間がかかる．このため，一度アクセスした有用な情報に再アクセスする際には，同じ手間をかけないようにすることが望ましい．

本論文ではユーザが参照した文書に再アクセスする際に，関連する文書に効率よくアクセスできるようにすることを目的として，検索語の類似度を用いた関連文書の検索手法を提案する．提案手法では，ユーザが有用と思われる文書リストを作成すると，システムは，ユーザが文書を検索する際に検索エンジンに入力した検索語をもとに，ユーザにより参照された文書と有用な文書の関連付けを行う．これにより，ユーザは，有用な文書とその関連文書に効率よく再アクセスできる．

2．関連文書の検索における問題点

文書から索引語を抽出する処理は索引付けと呼ばれる．索引付けの目的は，文書中から，その文書を特徴付ける語を漏れなく抽出することである．一般に何度も言及される語は重要であることから，文書中の語の出現頻度に基づく索引付けが用いられている．しかし，このような索引付けでは，個々のユーザが検索文書に対して持つ視点や概念が索引語の重みに反映されない．このため，あるユーザにとっては文書の内容を特徴付ける上で重要な語であっても，文書中の出現頻度が低い語の重みは小さく設定されてしまう．たとえば，ユーザの参照した文書が，【懐石料理】【京料理】【祇園祭】【天神祭】に関するものである場合を考える．ユーザが【懐石料理】と【京料理】の関連性が高く，【祇園祭】と【天神祭】の関連性が高いと考えていても，【懐石料理】と【天神祭】に関する文書で，「大阪」という語が頻出であり，【京料理】と【祇園祭】に関する文書で，「京都」という語が頻出であれば，【懐石料理】と【天神祭】，【京料理】と【祇園祭】が関連付けられてしまう．このような問題に対処するため，個々のユーザが持つ概念を表すシソーラスがあれば，文書に書かれている内容の意味を考慮した関連付けが行えるが，ユーザが手動で自らの概念を表す辞書を構築することや，関連文書の特徴語を目的別に全て書き留めておくことは，ユーザの手間が大きい．ユーザが自分にとって有用な文書やその関連文書に後で再びアクセスするための何らかの手がかりを残しておくことは必要であるが，そのためのユーザの手間が大きくなると，ユーザが本来行おうとする文書の検索や参照が円滑に行えなくなってしまう．

†京都大学大学院情報学研究科社会情報学専攻
‡財団法人 京都高度技術研究所
††立命館大学大学院法学研究科

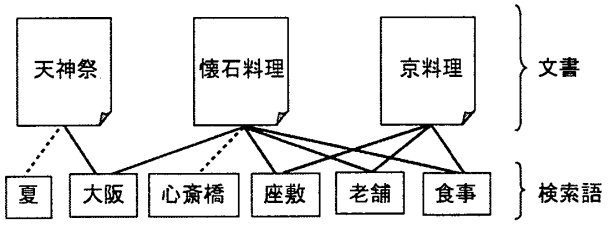

図1 類似文書の関連付け

3．検索語の類似度に基づく文書の関連付け

上に述べた問題を解決するために，文書中の頻出語ではなく，ユーザが検索エンジンに入力した検索語を索引語として，関連文書を検索する手法を提案する．ユーザが検索エンジンに入力する検索文（検索語の羅列）には，ユーザが検索しようとする文書の概念を的確に表す信頼性の高い特徴語が含まれる．このことから，文書中に含まれる全ての語を対象とするのではなく，文書の検索語のみを索引語として用いることにより，ユーザの持つ概念に即した関連文書の検索が行えると考えられる.なお，ここでは，AND演算子で結合されることを前提としてユーザにより検索エンジンの検索フィールドに空白区切りで入力される語をそのまま検索語として扱う．検索語に対する部分文字列の選択や複合語の生成，接辞処理などは行わない．

（1）類似文書の関連付け

検索文の類似度に基づく文書の関連付けについて，次に説明する．あるユーザが【懐石料理】に関する文書を「{大阪} AND {心斎橋} AND {座敷} AND {老舗} AND {食事}」，【京料理】に関する文書を「{京都} AND {鴨川} AND {座敷} AND {老舗} AND {食事}」，【祇園祭】に関する文書を「{京都} AND {夏} AND {山鉾} AND {厄除け} AND {浴衣}」，【天神祭】に関する文書を「{大阪} AND {夏} AND {風物詩} AND {みこし} AND {浴衣}」で検索した場合，たとえば，【懐石料理】と【京料理】では共通の検索語が 3 つあるのに対し，【懐石料理】と【天神祭】では共通の検索語が 1 つしかないため，【懐石料理】と【京料理】が関連付けられることとなる（図1）．このように，文書中の語の出現頻度ではなく，検索語の類似度に基づく関連付けを行うことで，このユーザに対しては，【懐石料理】と【京料理】，及び【天神祭】と【祇園祭】の関連付けが行われることとなる．

（2）文書間の類似度

共通の検索語を持つ関連文書が多数ある場合には，関連文書をユーザに提示する際に，ユーザにとって関連性の高いものの順に並べて提示することが望ましい．そのためには，文書の関連度を表す尺度が必要となる．

文書の関連度は，文書を索引語の重みベクトルで表現し，ベクトル間の類似度によって文書間の関連度を求める，ベクトル空間モデルに基づき計算することができる．

文書d_aとd_bの類似度を次のように定義する．

$$\sigma(d_a,d_b)=\frac{\sum_{i=1}^{T} w_i^a \cdot w_i^b}{\sqrt{\sum_{i=1}^{T}(w_i^a)^2}\times\sqrt{\sum_{i=1}^{T}(w_i^b)^2}}$$

ここで，w_i^nは文書d_nにおける検索語$t_i(i=1,...,T)$の重みである．w_i^nは，検索語t_iの網羅性$tf(t_i,d_n)$と特定性$iqf(t_i)$の積により，次のように定義する．

$$W_i^n = tf(t_i,d_n)\times iqf(t_i)$$

網羅性$tf(t_i,d_n)$は検索語t_iが文書d_nの検索文に含まれるかどうかを表し，検索語t_iが文書d_nの検索文に含まれる場合は 1，含まれない場合は 0 とする．特定性$iqf(t_i)$は，検索語t_iが文書をどの程度，特徴付けるか，すなわち，検索語t_iが全検索文の中のどれくらいの検索文に出現するかを表す尺度 IQF(Inverse Query Frequency)を表し，従来型の文書の関連付け手法における IDF(Inverse Document Frequency)に対応する．特定性の尺度 IQF を用いることにより，特定の少数の検索文にのみ出現する検索語の重みを大きくすることができる．特定性$iqf(t_i)$を次のように定義する．

$$iqf(t_i)=\log\frac{N}{qf(t_i)}+1$$

ここで，Nは検索文の総数であり，$qf(t_i)$は検索語t_iが出現する検索文の数である．

4．個人用アーカイブ管理支援

上記の文書関連付け手法を用いて開発した個人用 Web アーカイブ管理ツールについて，次に説明する．

（1）アーカイブの構築

個人用の Web アーカイブ構築ツールには，それ自身が単独で動作するアプリケーション型のものと，ブラウザと上位のサーバの間で動作するミドルウェア型のものがある．アプリケーション型のものは，アーカイブ作業をユーザが細かく指定しなければならない．一方，ミドルウェア型のものは，ユーザがアクセスしたものをキャッシングプロキシと同様の原理で自動的に保管するため，アーカイブ作業においてユーザの手間がかからないというメリットがある．本研究では，個人用のプロキシを用いることにより，ユーザが参照した文書を全てアーカイブに保存し，ユーザが検索エンジンに入力した検索語と閲覧した文書の履歴をプロファイル情報として保存する．

（2）ユーザによる有用文書リストの作成

ユーザは自分にとって有用な文書を発見した際に，後で再アクセスするための何らかの手がかりを残しておくことが必要である．再アクセスする際に，なぜその文書が有用であるかをユーザが思い出すことができなければならない．そこで，有用な文書の所在情報（Web ページの URL）と，文書に含まれる素材コンテンツ等にコメントを付加して規定のフォーマットで有用文書リストとして記述しておき，再アクセスのためのショートカットとして利用することとした．ユーザは有用文書のリストを作成する際に，我々の開発した専用のエディタを使用する．これにより，ユーザは，ブラウザの表示画面からのコピー&ペーストにより有用文書リストを作成することができ，ブックマーク保存程度の簡単な作業で有用な情報への再アクセスのための手がかりを残すことができる．

（3）有用文書の関連付け

ユーザが作成する有用文書リストと，ユーザが検索エンジンに入力する検索語をもとに，前節で述べた関連文書の検索を行い，関連文書を関連度の高い順でソートした関連文書リストを作成し，有用文書リストにリンクを設定する．これにより，ユーザが予め有用であると記述していない文書であっても，過去に参照した文書のうち有用文書に対する関連性が高い文書については効率よく再アクセスできるようになる．

5．関連研究

情報への効率的な再アクセスを目的として，アクセス履歴やブックマークの分析に基づくユーザ支援が提案されている([1][2][3][4][5])．我々のアプローチは，ユーザが作成する有用文書リストやアクセス履歴を利用するという点ではこれらの研究と同じであるが，文書の関連付けに Web サーチエンジンに対する検索語を利用する点が異なる．

6．むすび

ユーザにより入力される検索エンジンに対する検索語は，個々のユーザが持つ概念を的確に表す信頼性の高い特徴語である．検索語を用いた文書の関連付けを行うことで，ユーザの視点や概念に応じた有用性の高い文書の関連付けを行える．また本稿では，提案手法に基づき開発した個人用アーカイブ管理支援ツールについても述べた．複数のユーザ間で，知識を交換しあって，より効率のよい情報再アクセスを行うことが今後の課題である．

文　献

[1] L. Tauscher, and S. Greenberg, "How people revisit web pages: Empirical findings and implications for the design of history systems," International Journal of Human Computer Studies, 22, pp.549-562, 1997.

[2] R. M. Keller, S. Wolfe, J. R. Chen, J. L. Rabinowitz, and N. Mathe, "A Bookmarking Service for Organizing and Sharing URLs," in Proceedings of the 6th International World Wide Web Conference, Santa Clara, CA, April 1997.

[3] D. Abrams, R. Baecker, and M. H. Chignell, "Information Archiving with Bookmarks: Personal Web Space Construction and Organization," in Proceedings of the ACM Conference of Human Factors in Computing Systems (CHI '98), pp.41-48 , 1998

[4] W.S. Li, Q. Vu, D. Agrawal, Y. Hara, and H. Takano, "PowerBookmarks: A System for Personalizable Web Information Organization, Sharing, and Management," in Proceedings of the 8th International World Wide Web Conference, Toronto, Canada, May 1999

[5] S. Kaasten, and S. Greenberg, "Integrating Back, History and Bookmarks in Web Browsers," in Extended Abstracts of the ACM Conference of Human Factors in Computing Systems (CHI'01), pp.379-380, ACM Press, 2001