1G1-01

プロキシエージェントを用いた個人間協調情報検索システムの構築

Construction of the individual cooperation information retrieval system using the proxy agent

髙橋 亮太[*1] Ryota Takahashi　亀井 剛次[*2†] Koji Kamei　湯川 高志[*1] Takashi Yukawa　赤埴 淳一[*2] Jun-ichi Akahani

[*1] 長岡技術科学大学 Nagaoka University of Technology　[*2] 日本電信電話株式会社 NTT コミュニケーション科学研究所 NTT Communication Science Laboratories, NTT Corporation

This paper presents implementation of a collaborative concept-based information retrieval (IR) system. The authors have proposed a collaborative IR scheme for personal documents with a mechanism for adapting to divergence in word semantics. This scheme is targeted to IR on personal repositories. A personal repository is storage in a personal computer, and it stores personal documents which the user writes or downloads. To capture the user's word semantics, this scheme utilizes a concept base, which is a personal cooccurrence-based thesaurus that is constructed from the documents in the personal repository. In this paper, a system which constructs a personalized cooccurrence-based thesaurus from personally accumulated e-mail messages and access history of WWW pages is introduced and the implementation of the system is illustrated. The paper also reports experimental result which demonstrates that the system achieves to discern users' relation based on differences in the users' personality.

1. はじめに

個人が蓄積した情報を特定のコミュニティのメンバー間で共有もしくは交換し、コミュニティにおける情報流通に役立つようなシステムが必要とされている。筆者らはこれを満たすものとして、パーソナルレポジトリシステムを研究している。

パーソナルレポジトリシステムとは、個人の扱った情報をパーソナルレポジトリと呼ばれる記憶領域へ自動的に蓄積させていくシステムである。これを用いたシステムを個人がそれぞれ所有し、これらを連携させることで、他者が異なる情報源から得た情報の取得が可能となり、また、他者の価値観に基づいて整理された情報から新たな発見をすることも可能である。実行のためにはパーソナルレポジトリ相互の協調的な検索が必要であり、これを KV-P2P と名づけて研究している[湯川 2003]。協調検索を行う場合、個人間の関係を判別する必要があり、そのユーザの嗜好や個性といった要素を反映した検索結果を出せる物であることが必要とされる。また、ユーザの個人的な情報を扱うことからプライバシーについても充分考慮すべきである。

KV-P2P では、ユーザごとに異なるパーソナルレポジトリからそれぞれのパーソナル概念ベースを構築し、概念ベースの差異を数値で計測することが可能である。このシステムを用い、他者のパーソナルレポジトリを検索する技術や、さらに検索結果の正誤を判定できる技術が確立されれば、協調情報検索システムが実現されると期待されている。概念ベースとは、語に関する知識ベースであり、パーソナルレポジトリから構築されるパーソナル概念ベースは語の関係としてユーザの個性や嗜好を表現することになる。パーソナル概念ベースを比較することにより、個人間の関係を判別することが可能であると期待される。

個人が持つ文書の中でも、蓄積した電子メールは個人の嗜好や特性を良く表すと考えられる。また、実際には蓄積しないが、web ページの閲覧履歴も個性を表すだろう。これまでに、個々のユーザにおいて web ブラウザのブックマークに登録された web ページを蓄積文書と見立てた実験により、パーソナル概念ベースの個人適応性が確認されている。

しかしながら、実際に個人が蓄積した、あるいは閲覧した文書に基づく評価はされていなかった。そこで本稿では、電子メールと閲覧した web ページのすべてを対象とできる様なプロキシエージェントを実装した。これに基づいて、実データより概念ベースを構築して比較する実験を行い、KV-P2P が個人間の関係を判別できることを明らかにした。

具体的には個人間関係判別の実現について述べ、パーソナル概念ベース構築の評価結果を報告する。

連絡先:長岡技術科学大学, 〒940-2188 新潟県長岡市上富岡町 1603-1, Tel:0258-17-5143, bee@stn.nagaokaut.ac.jp
†現 NTT コミュニケーションズ株式会社 ソリューション事業部

2. パーソナルレポジトリと協調検索システム

2.1 パーソナルレポジトリシステム

(1) パーソナルレポジトリ

情報を蓄積し検索や操作を行うシステムの中でも、個人が持つ情報に特化したものをパーソナルレポジトリと呼ぶ。従来は、多くの人の持つ大量の情報を集積し一様に提供することを目指した、いわゆる図書館型の情報の蓄積・検索システムが多く開発されてきた。パーソナルレポジトリは、これとは反対に、個人が持つ情報を自身が活用することを目指したもので、書斎型のシステムと言える。

パーソナルレポジトリが図書館型システムと本質的に異なるのは、検索や操作に際して、それを所有する個人にシステムが適応することにある。このために、情報を蓄積する機構(レポジトリ)として半構造データベースが用いられ、情報そのものにさまざまなアノテーションを付加して蓄積できるようになっている。レポジトリに蓄積される典型的なアノテーションとしては、文献情報やメールのヘッダ情報などのメタデータが存在し、データの記述には RDF(Resource Description Framework)を用いている。また、検索に際しても, 所有者の意図を反映した検索を行う機構が用いられる。検索に際しては、次節に述べる概念ベースを用いた検索が好適であると考えている。

(2) KV-P2P

KV-P2P とは、日本語データから利用者ごとに異なる概念ベースを構築し、それを元に情報検索や個々の概念ベース間の差異の計測を可能にするものである。概念ベースは個人の蓄積した文書に基づいて構築され、個性や嗜好を反映しているが、KV-P2P ではこれを利用し次の機能を実現する。

- 概念ベースの差異に基づいて個人間の関係を判別する。
- 概念ベースの差異を等化して他ユーザの意図に沿った検索を行う。

個々のユーザが自身の興味を反映した概念ベースを持つ場合、他のユーザのレポジトリに対して情報検索を行うことには二つの効果が期待できる。一つは検索対象を広げる効果であり、他者の情報源を借りて自身の手元に届かない情報を取得することができる。これは公開の情報源に対する選好の差によるものだけでなく、メールや個人メモなどの非公開の情報源も含めて考えることができる。もう一つの効果は検索の幅を広げる効果であって,他者の概念を借りることによって,自身の判断基準では見過ごしてしまうような関連情報に気付くことが可能となる。

今回は、個々の送受信したメールと web ページの内容を自動的にパーソナルレポジトリへ蓄積するモジュールを統合させ、その日本語データから概念ベースを構築するという手法を用いた。

2.2 概念ベースとそれに基づく情報検索

(1) 概念ベース

概念ベースとは、辞書における単語の定義や文書中での単語の共起関係を利用して単語間の関係を定義し、類義語や関連語を含む文書の検索を可能とする、単語を 100～200 次元のベクトルで表したものである。

その構築方法は、はじめに、検索対象となる文書セットに含まれる文章を形態素解析する。次に、文中の単語の出現位置に基づき、図 2-1 に示すような近傍共起マトリックスを作成する。ここで近傍共起とは、ある単語 A と B が指定する範囲内に存在することを示す。そして、A と B が近傍共起している場合、A と B のマトリックスの交点の数に 1 を足し、語数×語数の近傍共起マトリックスを作成する。

	単語 1	単語 2	単語 3	…
単語 1	0	1	0	…
単語 2	0	0	0	…
単語 3	0	0	0	…
…	…	…	…	…

図 2-1 近傍共起マトリックス

その近傍共起マトリックスに特異値分解(SVD)を用いて、100～200 次元に圧縮し、これを個々の単語ベクトルとする。SVD とは, 任意の長方形の行列 A(N×M)を

$$A = ULV^T \qquad (1)$$

$$\begin{pmatrix} \\ A \\ \\ \end{pmatrix} = \begin{pmatrix} \\ U \\ \\ \end{pmatrix} \begin{pmatrix} \\ L \\ \\ \end{pmatrix} \begin{pmatrix} \\ V \\ \\ \end{pmatrix}^T$$

の形に分解する手法である。ここで、U は(N×M)の列正規直交行列、そして L は(M×M)の正規直交行列 である。また、L の対角成分は特異値(singular value)と呼ばれる。本稿では、パーソナルレポジトリに電子メールや web ページの内容を蓄積させて行き、これを概念ベース構築の対象文書セットとしている。図 2-2 は, SVD によって分解されたマトリックス U(N×M)の左側, 単語数×概念 100 のマトリックスを抽出したものである。このようなマトリックスを抽出する理由として, SVD 後のマトリックスの中で語の概念を最も多く含んでいる場所がマトリックス U(N×M)の左側 100 次元であることが挙げられる。よって、本稿では図 2-2 のようなマトリックスを概念ベースとして用いている。

	概念 1	概念 2	…	概念 100
単語 1	u_{11}	u_{12}	…	u_{1100}
単語 2	u_{21}	u_{22}	…	u_{2100}
単語 3	u_{31}	u_{32}	…	u_{3100}
…	…	…	…	…

図 2-2 概念ベース

(2) 概念ベースに基づく情報検索

文書ベクトルを生成する。そのために、文書ごとに形態素解析を行い、含まれる単語集合を取り出す。取り出した単語集合に含まれる単語それぞれに対し、概念ベースから対応するベクトルを得る。すべてのベクトルに出現回数に応じた重みをつけて総和し、その後、長さが 1 となるように正規化する。これを文書のベクトルとする。

検索の手順としては、問合せも文書ベクトルと同様にベクトルを生成する。文書ベクトルと問い合わせベクトルの内積を関連度として、すべての文書に対する関連度を計算し、関連度の大きい方から一定件数を検索結果とする。

この手法の利点は、問合せに指定した語そのものが文書に含まれていなくとも、概念的に類似した内容であれば高い関連度として検索されることと、パーソナルレポジトリに適用した場合には、概念ベースがユーザの嗜好や志向を反映することになるので、個人適応した検索が実現できることである。

(3) 概念ベースに基づく個人関係判別

パーソナルレポジトリにおける概念ベースは上述の通り、個人の嗜好や志向を反映している。個人間で概念ベースを比較することで、嗜好や志向に基づいた個人間の関係を判別することが可能であると期待される。具体的には、概念ベースの差異の尺度を以下のように定義する。S と T を概念ベースとし、それらに含まれる語の数は同じであるとする。S における語 v と語 w の間の類似度を sim^S_{vw} と書く。このような S における語 v と語 w 語の間の類似度は、

$$sim^S_{vw} = \frac{\vec{v}_S \cdot \vec{w}_S}{|\vec{v}_S| \times |\vec{w}_S|} \qquad (2)$$

である。

語 v と語 w の、概念ベース S と概念ベース T における類似度の差は、

$$d_{vw} = \left| sim^S_{vw} - sim^T_{vw} \right| \qquad (3)$$

である。

概念ベース全体について差の値を集計する。M を概念ベース中の語の数とする。上記の式をすべての単語のペアについて計算し、m^2 個の値が得られる。求める値は、

$$diff_{S,T} = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\left(sim_{vw}^{S} - sim_{vw}^{T}\right)^2}{m^2}} \qquad (4)$$

となる。

3. データ入力モジュールの実装

パーソナルレポジトリにおける個人間の関係判別には、

1.個人のデータをパーソナルレポジトリに入力または変換する機能

2.蓄積した文書とメタデータを扱うパーソナルレポジトリ管理機能

3.概念ベースの構築機能

4.複数の個人の概念ベースを比較する機能

が必要であるが、2～4は既に実装されている[Yoshida 2003, 亀井 2003]。よって今回、1のデータ入力モジュールを新たに実装した。ここでは本モジュールの実装の方針と詳細を述べる。

電子メールメッセージと閲覧した web ページのパーソナルレポジトリへの入力は、ユーザに特別な操作を強要することなく、また、意識させずに出来ることが望ましい。

そこで、

1.電子メールについては、IMAP サーバと通信をインターセプトしてメッセージを取り出し、パーソナルレポジトリへ入力する Mail Monitor として、

2.web ページについては、proxy サーバとして振る舞いつつ閲覧された web ページをパーソナルレポジトリにも入力するものとして、それぞれ実装した。

図 3 にパーソナルレポジトリシステムの概要を示す。通常、IMAP サーバを利用する場合、各自のメールクライアント が IMAP サーバに接続し、メールの送受信を行う。今回は MailMonitor に IMAP サーバの挙動を監視させ、送受信したメールの内容を逐一パーソナルレポジトリへ保存させている。同様に、web ブラウザで web ページを閲覧する場合においても、プロキシを通して閲覧した内容をパーソナルレポジトリへ蓄積させてゆく。IMAP サーバとの通信は JavaMail を用いて実装した。

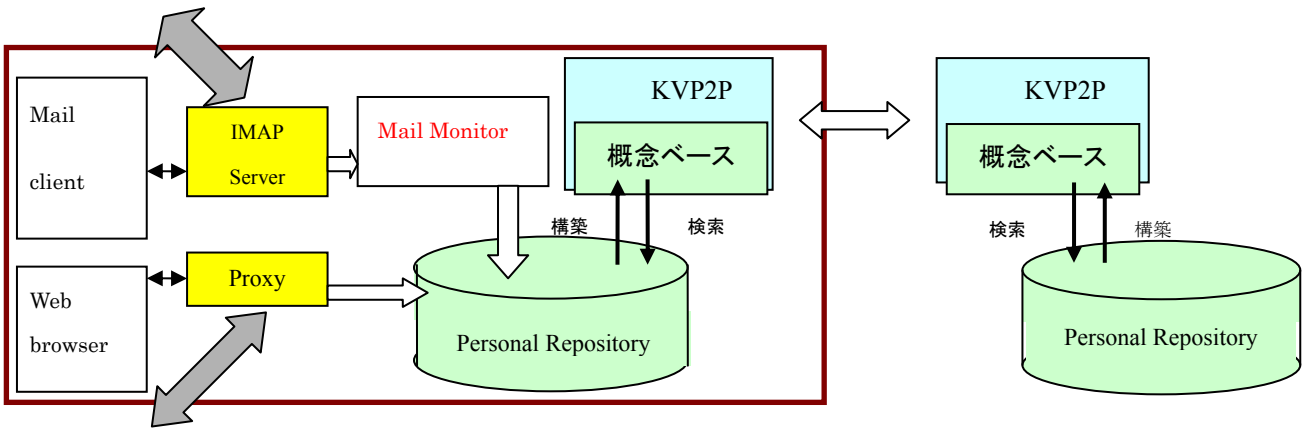

図 3 パーソナルレポジトリシステムの概要

KV-P2P はパーソナルレポジトリへの文書の蓄積を監視しており、処理可能な文書、ここでは Content-Type が text/html および text/plain で、chaset が日本語(ISO-2022-JP, Shift-JIS, SJIS, EUC-JP)であるものが保存されたときにその文書を解析し、検索対象として登録する。また蓄積された文書を元に KV-P2P を用いて概念ベースを作成する。作成された概念ベースは他者の概念ベースと比較し、概念ベース間の差を調べることが可能である。

メールクライアントに対してではなく、あくまで IMAP サーバを対象にデータ収集を行うことで、ユーザが使ってきたツールを変更しなくてよい透過型のプログラムとして実現できている。

ユーザが送信するメールについても、バックアップを IMAP サーバに保存する設定にすることで取得可能としている。

4. 実データに基づく個人関係判別の評価

4.1 実験方法

概念ベースに基づいた個人間関係判別の有効性を確認するために、5 人の被験者による実験を行った。

一ヶ月分の電子メールとこの間に閲覧した web ページを、前述のデータ入力モジュールによってパーソナルレポジトリに収集し、2.2 節で述べた差異の尺度により概念ベースを比較した。

5 人の被験者(A,B,C,D,E と表記する)の主な特徴は次の通りである。

- A は FreeBSD 関連のメーリングリストを講読
- C は他の 4 人が購読しているパーソナルレポジトリ関連のメーリングリストの購読をしていないが、linux 関連のメーリングリストを講読している
- E は実験開始の一月前から合流した新規メンバー
- 5 つのプロジェクトが存在し、それに携わっているメンバーの関係は以下の通り
  - プロジェクト 1(A-B)
  - プロジェクト 2(A-D)
  - プロジェクト 3(C-D)
  - プロジェクト 4(D-C)、ただし C はあまり深くは関わっていない
  - プロジェクト 5(B-E)

一つのプロジェクトに二人が関わっているが、前方に挙げた人物が主導的立場にある。

4.2 実験結果

本実験で個人の手元に出現した単語数は、4000～20000 程度の幅があり、そのうち最大 5000 を取って処理対象とした。5000 語のうちの大体 60%程度が重なっていたので、二人の総

合単語数を取ると 6500 から 7000 語になった。これらのレポジトリ間で存在する差異を評価したものを図 4-1、図 4-2 に示す。図において値が小さいほど概念ベース間の差異が小さく、概念ベース間の差異が小さいほど、嗜好や志向が近いことを表しているはずである。

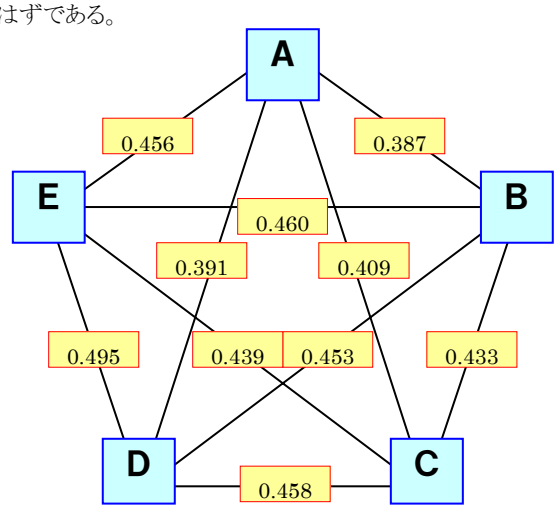

図 4-1　全体を比較した概念ベース間の差

図 4-1 は全体を比較した概念ベース間の差であり、二者の持つ全単語を元に概念ベースの差を求めたものである。それに対し図 4-2 は共通語彙のみを比較した概念ベース間の差であり、二者の持つ共通の単語のみを元に概念ベースの差を求めたものである。結果を見ると、A と B、A と D は図 4-1、図 4-2 ともに差異が相対的に小さく、関係が深いと判別される。A と C は共通語彙における差異が小さく、嗜好や志向が近いと分かる。E は他メンバーの誰とも差異が比較的大きく、これは他のメンバーとの関係が浅いと判別される。

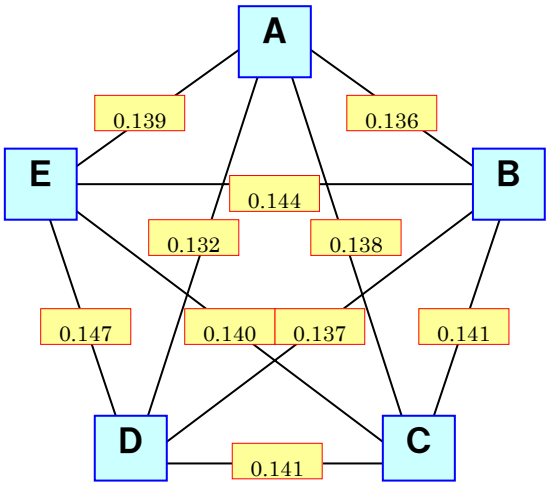

図 4-2　共通語彙のみを比較した概念ベース間の差

4.3 考察

A については、B、または D と共通のプロジェクトに取り組んでいる。図 4-1 と図 4-2 を見ると、A-B と A-D はどちらも A-C と A-E 両方より差が少なく、図 4-2 では数値が 0.400 を切っている。この数値は周囲と比較し低くなっていることが分かる。

B の場合は、図 4-2 によると差が大きい順に E,C,D,A となっている。E が新規メンバーであること、A とはプロジェクト 1 関連で蓄積される情報に共通性がある。

他の 4 人が購読しているメーリングリストを唯一購読していない C は、図 4-2 によると B,D,E の三人に対して 0.140～0.141 の範囲であるが、A に対しては 0.138 と他よりも小さい値となっている。A-C の親和性は、A と C が購読しているメーリングリストの結果が現れたものであると考えられる。

D については、図 4-1、図 4-2 ともに A,B,C,E の順に小さい値となっている。A-D はプロジェクト 2 の関連より小さく、D-E は交流の短さから大きい値となったと判断できる。

新規メンバーの E が関わっている組み合わせでは、概念ベース間の差が他よりも大きくなっている。これは、新規ゆえにその事が理由で誰からも離れた感じがあると見られる。

これより、被験者間で同一のプロジェクトを手掛けている等、人間関係や話題性において関連の深い場合は概念ベースの差を表す数値が小さい傾向にあると分かり、逆に新規メンバーといった立場のように他メンバーとの馴染みが薄い場合は、概念ベースの差を表す数値が他と比較して高い傾向にあるという結果が出た。この結果は、本システムにおける概念ベースの差異の指標が個人間の関係を良く表しているといえる。

5. おわりに

本稿では、個人間協調検索機構 KV-P2P について述べ、それが持つ概念データベース構築機能を用いて個人間の関係を判別するシステムの実装と実データによる評価を行った。電子メールと web ページを、ユーザに意識させることなく、パーソナルレポジトリに蓄積できるよう、プロキシエージェントとしてのデータ入力モジュールを実装した。5 人の被験者による実データを用いた実験により、話題や嗜好・志向の共通性が高いと、概念ベースの差異が小さくなり、これによる個人間の関係判別が可能であるとの示唆を得た。

今回の結果より、個人の情報を収集することによる技術向上が本実験により確認されたが、それは近い将来、この技術が一般に浸透する場合のことも視野に含み、個人の情報を得ることによる精度向上とどこまで情報を収集すべきかといったプライバシー問題の線引きを考えることも今後の課題の一つであると言える。

参考文献

[湯川 2003] 湯川 高志, 吉田 仙, 桑原 和宏: 協調パーソナル・レポジトリ・システムとその情報検索機構, 信学技報, vol.102, no. 602, pp.37-42, 2003.

[阿部 2003] 阿部 仁志, 湯川 高志: 類似概念判別に基づく情報検索システムの検索性能の評価, 人工知能学会全国大会(第 17 回), 1C4-04, 2003.

[亀井 2003] 亀井 剛次, 湯川 高志, 吉田 仙, 桑原 和宏: パーソナルレポジトリ間の協調情報検索, JAWS2003, pp.329-336, 2003.

[湯川 2001] 湯川 高志, 吉田 仙, 桑原 和宏: パーソナル・レポジトリに対するピア・ツー・ピア型協調情報検索機構の提案, 信学技報, vol.101, no.420, pp.9-16, 2001.

[Yoshida 2003] Yoshida, S., Yukawa, T., and Kuwabara, K.: Constructing and Examining Personalized Cooccurrence-based Thesauri on Web Pages, in Proceedings of the Twelfth International World Wide Web Conference(WWW2003), 2003.