D-008

相関ルールに基づく文書検索システム

A Document Search System Based on Association Rules

竹下　日出男† 大石 哲也†† 長谷川 隆三††† 藤田 博††† 越村 三幸†††

† 株式会社日立製作所

†† 九州大学大学院システム情報科学府

††† 九州大学大学院システム情報科学研究院

1 序論

文書データの蓄積と検索においてデータベースが用いられるのが一般的である．実際，ニュース記事や学術論文などの文書データベースが存在する．また，これらのデータベースから快適かつ手軽にユーザの求める情報を得るための手法に関して多くの研究がなされている．

検索の際，ほとんどのシステムがユーザに検索語(クエリ)を入力させ，対象となる文書中の語との一致を見ることで，スピーディな情報検索を可能としている．このとき，クエリの選択次第で検索結果の良し悪しが左右されるわけだが，必ずしも最適なクエリをユーザが入力できるとは限らず，結果的にユーザは，何度もクエリを考え直すことになる．

また，情報検索を行う動機は様々考えられる．現在はインターネットを通じて刻々と新たな情報が得られるため，今注目している記事に関係する情報や，今見ている論文に類似した論文を探したいといった「手元のテキストデータに類似した情報を得たい」という欲求は比較的頻繁に生じる．そのような場合，手元の文書からユーザの関心となる部分を抽出し([3], [4])，それを検索に利用する事が有用と考えられる．

そこで，本研究ではユーザが現在関心を持っているテキストデータを分析することにより，それに関連のある文書をデータベースから検索し，ユーザに提示するシステムを提案する．文書間の関連を求める手法については，相関ルールの概念を用いる．相関ルールは事象間の同時性や関連性を表現する手法の一つである．他方，テキストマイニングにおいては，文書データベースに出現する単語の統計的傾向から，単語間の関連性を見出しているものもある([1])．

2 節では本論文で用いている相関ルールの考え方について簡単に述べる．3 節では提案システムの概要について述べ，システムを構成する各部分について説明する．4 節でこのシステムを用いて行った検証実験について述べ，5 節で今後の課題に触れ本文をまとめる．

2 相関ルール

ある事象が発生すると別の事象が発生するといったような，同時性や関連性の強い事象の組み合わせ，あるいはそうした強い事象間の関係のことを相関ルールと呼ぶ([7])．

一般に相関ルールは X, Y を事象の集合として，

$$X \Rightarrow Y \tag{1}$$

と言う形式で記述される．ここで X を前提部，Y を結論部と呼ぶ．

相関ルールの形式的な定義は次のようになる．$I = \{i_1, i_2, \ldots, i_m\}$ をアイテムの集合とする．トランザクション T はアイテムの集合であり(すなわち，$T \subseteq I$)，データベース D はトランザクション T の集合である．

データベース D 中のアイテム集合 X を含むトランザクションのうち，アイテム集合 Y を含むものの割合が c% であるとき，「相関ルール $X \Rightarrow Y$ は D において c% の確信度(confidence)で成立している」と言う．

また，D 中のアイテム集合 $X \cup Y$ を含むトランザクションの全てのトランザクションに対する割合が s% であるとき，「相関ルール $X \Rightarrow Y$ は D において，s% のサポート(support)を持つ」と言う．

確信度は，相関ルールの確からしさを表す指標であり，サポートは相関ルールの汎用性を表す指標である．価値のある相関ルールを探し出すときには，2 つの指標がそれぞれある程度以上大きいことが望ましい．

3 提案システム

本節では，ユーザの関心のある文書を元に，データベースに問い合わせを行い関連のある文書を提示するシステムに関して説明する．

3.1 システムの概要

システムは大きく 2 つの部分から成り立っている．文書データ群を入力とし，データベースを構築する

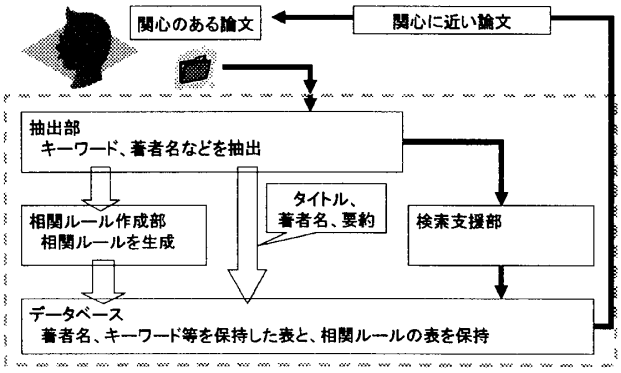

図 1: システムの概要

ルールデータベース構築部．ユーザからの文書を入力とし，データベースに問い合わせを行う関連文書検索部である．

3.2 データ入力

文書から重要語を抽出し，トランザクション化する抽出部と，相関ルールデータベース構築エンジンからなる．相関ルールデータベースの構築 ([2], [5]) は，対象とする文書データベースが更新されたときに行われる．ルールデータベース構築の流れを図 1 の白抜き矢印で示した．

1. 対象とする文書データベース上の文書を分析し，必要なデータと特徴的な語を抽出する．抽出した特徴的な語の集合を相関ルールを適用する際の 1 つのトランザクションとする．この操作を全ての入力文書群に対して行い，自動的に割り振られる ID(TID) を識別子としたトランザクション群を構築する．

2. 構築されたトランザクション群に相関ルール生成アルゴリズムを適用し，表 1 のような相関ルールデータベースを構築する．

3.3 データ検索

ユーザから元となる文書を入力して貰い，入力文書の解析結果を元に関連文書をデータベースの中から検索する．検索には予め構築している相関ルールデータベースを利用する．データベース検索の流れを図 1 の実線矢印で示した．

1. ユーザは現在関心がある文書をシステムに提示する．これを受けてシステムは提示された文書を分析する．すなわち，特徴的な語を抽出し，1 つのトランザクションを構築する．この際，特徴語抽出のときに，3.2 節の段階で抽出されている相関ルールの前提部に含まれる語を優先的に抽出するようにする．後のマッチングが効率的に行われるようにするためである．

表 1: 相関ルールの例

{ Web}	⇒	{ 検索 }
{ 検索, 情報 }	⇒	{ 手法 }

2. 提示文書から構築されたトランザクションを元に，データベースに対する問い合わせを作成する．事前に作成した相関ルールを利用する．提示文書の特徴的な語を前提部に含むような相関ルールがある場合，その相関ルールの結論部を問い合わせに追加する．具体的には，提示文書の特徴的な語に「Web」を含んでいた場合，表 1 の 1 行目のルールを用いて，「検索」も問い合わせに追加する．

3. 相関ルールを参考に，問い合わせとなるクエリの組み合わせの順をソートする．ソートされた順のクエリの組み合わせでデータベースに問い合わせし，結果をユーザに提示する．

4 実験および評価

本節では，本システムを利用した実験を行い，その結果についての評価を行う．まず，実験の概要について説明し，その後に具体例を用いて実験結果の評価を行う．

4.1 実験の概要

実験は本システムを実際に用いて提示される文書の評価について行う．

実験では，本システムと全文検索システム Namazu [6] の比較を行った．情報科学分野の論文約 550 件に対し，予めそれぞれの元となる論文 (ユーザの関心があると仮定した論文) に対し目標となる論文 (実際に元の論文と関連のあるとした論文) を定め，どれだけの目標論文を提案出来るかを調査した．

今回の実験では

- ユーザが関心を持っている文章データ（文書 A)

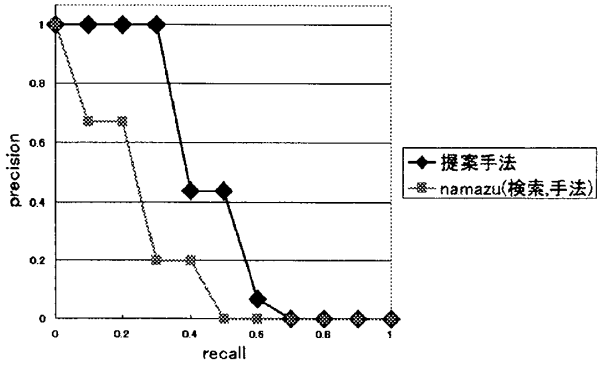

図 2:「関連文書の提案」に関する記事を入力とする Recall-Precision グラフ

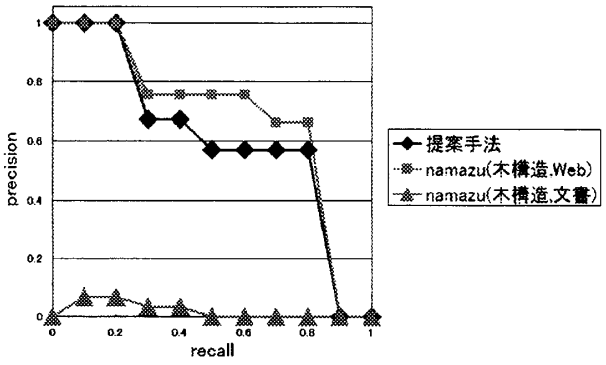

図 3:「半構造化文書からの木構造抽出」に関する記事を入力とする Recall-Precision グラフ

- 文書 A に関連のある文書群 P

を予め定めておくことにする.

以降の実験の中では，提案システムに対しては，文書 A を入力とし，全文検索システム Namazu に対しては，文書 A から連想される幾つかの単語を検索語として実験を行った．今回は文書 A のサンプルとして情報検索分野の論文から，以下の 2 つの論文を利用した.

- 「関連文書の提案」について述べた論文，文献 [8]
- 「半構造化文書からの木構造抽出」について述べた論文，文献 [9] を和訳したもの

4.2 実験結果の評価

結果の評価には，情報検索の分野で広く用いられている再現率 (recall) と適合率 (precision)，および 11 点平均適合率の考え方を用いる.

再現率は検索漏れの少なさを示す尺度であり，

$$\text{再現率} = \frac{\text{検索された文書中の適合文書数}}{\text{全文書中の適合文書の数}} \quad (2)$$

で表される.

適合率は検索のノイズの少なさを示す尺度であり，

$$\text{適合率} = \frac{\text{検索された文書中の適合文書の数}}{\text{検索された文書の数}} \quad (3)$$

で表される.

4.2.1 全体的な評価

全文検索を用いた実験で，図 2 と図 3 に示すような結果が得られた.

図 2 は「関連文書の提案」について述べた論文に対する結果である．全文検索よりも提案手法の方が高い検索効率を示した．この例の場合，全文検索に対するクエリを変更しても，提案手法より高い検索効率を示すものはなかった.

また，図 3 は「半構造化された文書から木構造を抽出する技術」について述べた論文に対する結果である．全文検索に対するクエリとして，この論文の要点と思われた「木構造，文書」を入力とした場合は，ほとんど目的とする文書に到達する事が出来なかった．しかし一方で,「木構造，Web」をクエリとして検索を行ったところ，高い検索効率を示した.

このように，クエリを用いた全文検索は，ユーザの検索能力 (関連文書を検索する際に適切なクエリを選択する能力) や，文書に含まれる文字の組み合わせに大きく依存する．何度も試行すれば，よりよい結果が得られる可能性はあるとは言え，クエリを何度も考え，そのたびに検索結果を確認するのは面倒であろう.

提案手法は，入力文書を分析した結果得られた関連性の高いであろうクエリの幾つかを自動的に問い合わせ，その結果を足し合わせる事で，ユーザにとっては 1 度の試行で，平均的に高い結果を得る事に成功している.

4.2.2 検索結果の評価

Recall-Precision グラフ上での提案システムの描く曲線からは次の 2 つのことが読み取れる.

- 再現率の低い部分での適合率が高いこと．
- 逆に，再現率の高い部分では適合率が低くなっていること．

前者は「グラフの上位での目的文書提案率が高い」ことを示す．一般的なユーザはあまり下位に提案された文書までは目を通さない傾向にあるので，上位での目的文書提案率が比較的高い傾向にある本システムは有効であると言える．

後者については，大きな問題がある．本システムでは，多くの場合，目的文書とした文書の内数件を，発見できないことがある．これは，システムが単語をクエリとした文字列一致型の検索であり，データベース内の文書に評価値を与え，その順位を並び替えるものではないことに起因している．データベース内の全ての文書に順位付けをして提案するものではないため，検索漏れが生じる可能性があるのである．

実験結果を見ると，単純な全文検索と比較すれば，全体的により多くの目的文書を発見している．しかしながら，目的文書の検索漏れをさらに減らすことは今後の課題である．

5 結論

本研究では，文書データベースに対して利用者が現在関心を持っている文書・記事データを提示する事により，それに関連する文書を自動的に検索するシステムを提案した．本システムは相関ルールの概念を問い合わせ単語の洗練に用いる点で新しく，次の 2 つの点が大きな特徴である．

1 つは，従来の検索手法がキーワードをクエリとしたものであるのに対して，1 つの文書全体をクエリとした手法であるという点である．これにより，ユーザが最適な検索キーワードを考える手間は小さくなり，また多数の関連語を検索のヒントとして用いることを可能にした．実験の結果によると，本手法は一般的な全文検索と同等かそれ以上の精度を持つことが示された．単純なキーワード検索では多くの場合，検索の多数の再試行が必要となるのに対し，一度で所望の検索結果を得られる本手法はユーザの負担を著しく軽減できる．

もう 1 つは，検索結果の上位においての適合率が高く，最初に表示される 20 件に目的文書が多く含まれることである．一般的なユーザは検索結果の下位までは確認しないと思われるので，目的とする文書が上位に集積できるだけでも，結果に対するユーザの満足度は著しく高まる．

残された課題もある．提案システムは，相関ルールデータベースを作るときに重要と思われる特徴語を抽出し，文書を特徴付けるアイテム集合としている．このアイテム集合が大き過ぎれば，データベース作成に時間が掛かるようになり，また，不必要な相関ルールを作り出すノイズ的な単語も多く含まれてくる．適切なアイテム集合の大きさの設定は，どのような語を抽出し相関ルールを作り出すかと合わせて重要な課題である．

参考文献

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining Associations between Sets of Items in Massive Databases, *Proc. of the ACM-SIGMOD Int'l Conference on Management of Data*, pp.207-216, 1993.

[2] M. Houtsma and A. Swami. Set-Oriented Mining for Association Rules in Relational Database, *Proc. of the 11th Int'l Conference on Data Enginnering*, pp.25-33, 1995.

[3] 「専門用語自動抽出システム」，東京大学中川研究室・横浜国立大学森研究室，http://gensen. dl. itc. u-tokyo. ac. jp/

[4] 中川裕志，森辰則，湯本紘彰: "出現頻度と連接頻度に基づく専門用語抽出"，自然言語処理，Vol.10 No.1, pp. 27 - 45, 2003 年 1 月

[5] 佐伯 敏章，新谷 隆彦，茂木 和彦，田村 孝之，喜連川 優，概念階層を考慮した相関ルールマイニングの関係データベース管理システム上での実現，情報処理学会 研究報告 Vol.98, No.57, (データベースシステム 116-17)，pp.127-134, 1998

[6] 全文検索システム「Namazu」，http://www.Namazu.org/

[7] 福田剛志，森本康彦，徳山豪，"データマイニング"，共立出版,2001.

[8] 竹下日出男，"相関ルールに基づく文書検索システム"，九州大学修士論文，2007.

[9] D.C. Reis, P. B. Golgher, A. S. Silva, and A. H. F. Laender. Automatic web news extraction using tree edit distance. *WWW2004*, pp.502-511, 2004.