複数ニュースサイトのいもづる式検索エンジン「いもなび」

山田 剛一　　大熊 耕平　　増田 英孝　　中川 裕志 †
東京電機大学 工学部　　† 東京大学 / 社会技術研究システム
{yamada@,ohkuma@cdl.,masuda@}jim.dendai.ac.jp, nakagawa@dl.itc.u-tokyo.ac.jp

1　はじめに

「いつでも、どこでも、誰でもインターネット」な現在、多くの人にとって Web 空間は最も身近な情報源であり、その入り口である Web 検索エンジンは情報収集のための必須の存在となっている。

Google をはじめとする Web 検索エンジンでは、単語をいくつか入力すると、Web ページ群を順位つきで提示するという形態をとっている。Web 検索エンジンは Web ページを探すための手段として提供されており、ページ一覧を結果として表示するのは当然であるが、実際にはユーザが Web 検索エンジンに言葉を入力するとき、その言葉ズバリのページを見つけたいとは限らない。例えばあるユーザが「地震」という言葉を入力したとき、「地震」ズバリのページ、例えば地震とはどういうものか、というページを読みたいとは限らない。『いま「地震」という言葉で検索すると、どんなページが出てくるだろう?』という興味で検索をすることもある。この場合、「地震」自体を知りたいのではなく、「地震」に関連するトピックが見えることを期待しているのである。

我々が開発している検索エンジン「いもなび」では、関連するトピックを知りたいという欲求に応えるよう、検索結果表示に単語でトピックを提示する仕組みを組み込んだ (図 1)。また、これらの単語を取捨選択し、その単語で再検索することが可能になっており、さらなる関連情報を得ることができる。この再検索の仕組みにより、ユーザは興味の趣くまま、いもづる式に情報を得ていくことができる。通常の検索エンジンが提供する単発の検索に対し、この連続的な検索こそが、いもづる式検索エンジン「いもなび」の最大の特徴であるナビゲーション機能である。

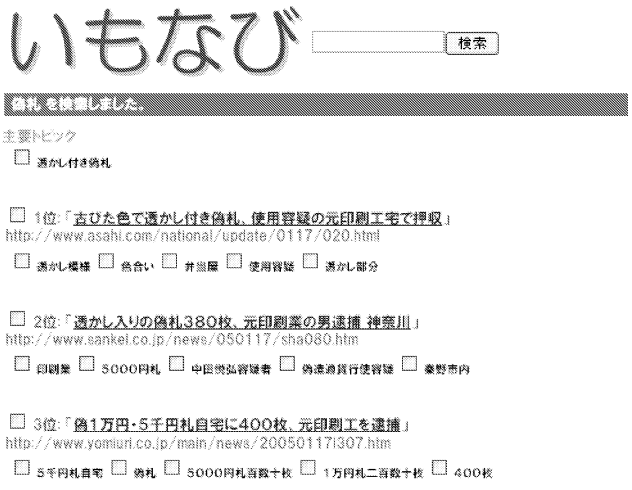

図 1: 「いもなび」の検索画面

「いもなび」では検索対象を Web で公開されているニュース記事に絞り、類似記事を収集してその差分を求めることにより、主要なトピックとは別の、周辺のトピックを得ている。この手法の考え方については次の 2 節で詳しく述べる。

図 2〜図 4 でナビゲーションの流れの例を示す。図 2 は最初に「コンピュータ」「ウイルス」という単語で検索を行った場合の例である。検索結果にはコンピュータウイルスの記事が得られているのがわかるが、内容に異なりがあるため、それぞれの記事にその差異である単語群が提示されている。ここで次の検索を行うために「感染報告」「累計」という単語を選択し検索を行なう。この検索結果を図 3 に示す。検索の結果としてエイズ関連の記事が得られていることがわかる。コンピュータウイルスの話題から、病気のウイルスの話に変わっている。次に「報告書」「困難」という単語を選択して検索を行なう。この検索結果を図 4 に示す。検索結果としては三菱ふそうのハブの欠陥についての記事が得られた。

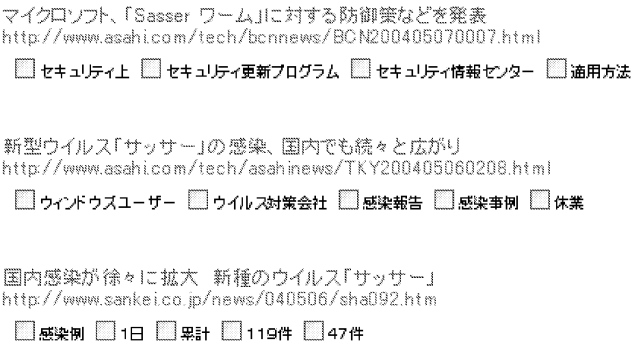

図 2: 「コンピュータ」「ウイルス」の検索結果

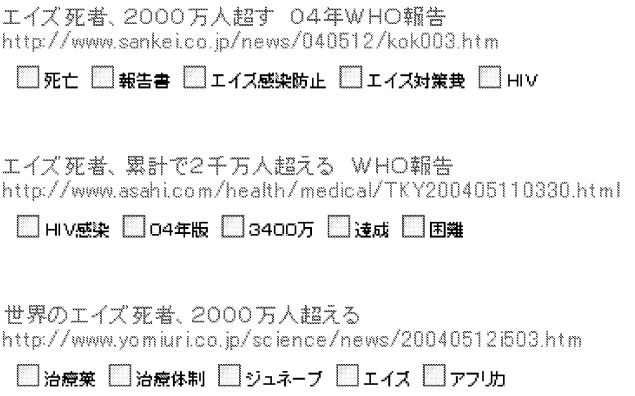

図 3: 「感染報告」「累計」の検索結果

この例では「コンピュータウイルス」→「エイズ問題」→「三菱ふそうの欠陥の問題」とトピックが移り変わっていった。このように、検索した結果の記事の差分を選択していくことで、ユーザは新たなトピックへとナビゲートされていく。

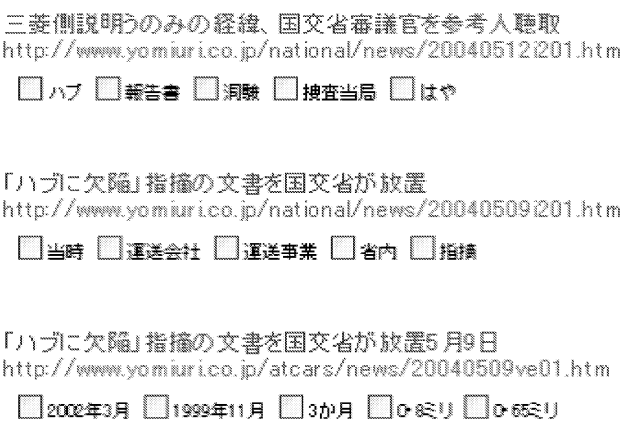

図 4: 「報告書」「困難」の検索結果

2 基本コンセプト - 差分の提示

類似する記事群には、いろいろな観点からの情報・見方などが多数示されている。この類似記事を用いることで、ある話題について単独の記事ではカバーしていなかったテーマへの糸口が見つかりやすくなる。

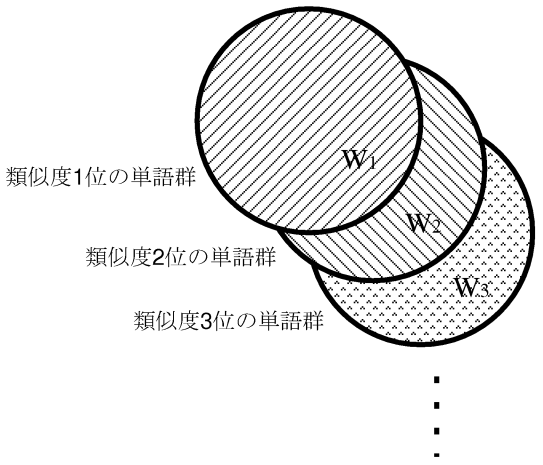

図 5: 類似記事群における単語の重なり

類似記事群はおよそ図 5 のように内容がオーバーラップしている。検索質問と最も類似度が高い記事はユーザが全文を読むことを想定する。ここで、1 番目の記事で扱っていない内容をいもづる式に手繰るという局面を想定してみる。同じような問題設定は多文書自動要約に見られる [1]。多文書自動要約では、MMR(Maximal Marginal Relevancy) という考え方が使われる。すなわち、元の質問に類似していることと、既に選択した記事に類似していないことの両者を加味した基準によって記事を選択する。我々の目的では、いもづる式ナビゲーションの性質上、既に選択した記事に類似していないことにのみ重きをおくことになる。したがって、問題は類似度が 2 番目以降の記事に出ている内容のうち 1 番目の記事と重複しないものをどのようにブラウザ上に提示するかである。要約における MMR の場合、表示は記事単位であった。しかし、ここでは記事全部を提示してしまうと、いたずらに利用者に負担を強いる。そこで、2 番目以降、例えば i 番目に類似した記事内容を提示する方法は、「i 番目の記事の内容のうち、既に選択した $i-1$ 番目までの記事に出現していない内容を表す部分を表示する」という方法をとる。しかし、このような部分に対応する文の集合を探し出すことは、意味理解に近いことが必要であり、現実的ではない。より軽い処理で実現でき、かつ利用者にも indicative な情報を提示できるという観点からは単語を単位とする表示が現実的である。そこで本システムでは、図 5 の重ならない部分を単語の集合とみなすことにした。i 位の記事と上位記事との差分であるこの単語集合を P_i、i 位の記事に含まれるすべての名詞の集合を W_i とすると、P_i を求める式は式 (1) となる。

$$P_i = W_i \cap \left(\bigcap_{k=1}^{i-1} \overline{W_k} \right) \quad (1)$$

$$(\text{ただし } P_1 = W_1 \cap \overline{W_2})$$

本論文では、この単語集合 P_i を特徴語と呼ぶことにする。

単語を提示するもうひとつの利点は、単語には TFIDF などの方法で重要度がつけられ、その重要度の順に表示するというコンパクトな表示ができる点である。

よって提案するシステムでは、収集した記事の見出しとその記事の特徴語を提示し、この特徴語を選択し再び検索を行うことでユーザをナビゲートする。

3 「いもなび」の実装

「いもなび」は通常の Web 検索エンジンと同様、Web ページを収集しインデックスを作成するクローラー/インデクサと、ユーザの検索要求を受け付け検索結果を提示する検索エンジンとからなる。

3.1 クローラー/インデクサ

「いもなび」は新聞社が公開しているニュース記事を収集対象としている。新聞社のサイトに普遍的な特徴として、記事一覧のページが存在し、そのページから記事のページへのリンクのアンカー文字列として、そのリンク先の記事の見出しが用いられている、というものがある。この性質を用いて、リンク先が記事ページか否かを判断し、不要なリンクを追わないようにしている。この方法で、主要新聞社 5 社のサイト (朝日新聞、読売新聞、毎日新聞、日経新聞、産経新聞) の記事を収集している。

現在のシステムでは、インデックスは単純語に対してのみ作成している。ただし、記事間の差分を求め表示する際には複合語単位で処理するため、複合語としての出現の情報もデータベースに格納する。

3.2 検索エンジン

3.2.1 記事の並び替えによる類似記事の取得

一般の Web 検索エンジンではリンク構造解析によるページのランキングが主流であるが、「いもなび」では対象がニュース記事であり、その内容の類似性に着目することから、ベクトル空間法によるランキングを行っている。ベクトルの要素は単語の TFIDF である。検索要求ベクトルと記事文書ベクトルの類似度を求め、その類似度の高い順に記事を提示するのが通常であるが、「いもなび」の検索エンジンではもう一段階の処理を追加している。

「いもなび」では、類似記事間の差分を求めることにより周辺トピックを提示する。そのため、検索結果の上位の記事は類似していることが望ましい。しかし、記事を検索要求との類似度の順に並べると、記事間の類似度が高くなるとは限らず、上位に類似記事が集まらない。そこで、検索要求との類似度が最も高い記事を新たな検索要求に見立て、その記事との類似度が高い順に記事を並べ替える。

検索要求のベクトルを V_{key}、記事のベクトルを $V_{URL_i}(i = 1,2,\cdots,N)$ とする。ただし、N は検索要求にマッチした記事数である。類似度 Sim_i は、V_{key} と V_{URL_i} との cosine 値とする。この式を式 (2) に示す。

$$Sim_i = \frac{V_{key} \cdot V_{URL_i}}{|V_{key}||V_{URL_i}|} \tag{2}$$

$V_{key} \cdot V_{URL_i}$ は V_{key} と V_{URL_i} の内積、また $|V_{key}|$ と $|V_{URL_i}|$ はそれぞれのベクトルの大きさである。これにより Sim_i が最も大きくなる記事 A_1 を求める。次に A_1 に関連する記事を選び出す。この式を式 (3) に示す。

$$Sim_j = \frac{V_{A_1} \cdot V_{URL_j}}{|V_{A_1}||V_{URL_j}|} \tag{3}$$

この式により得られた Sim_j の値の大きい順に記事を並べる。このようにすることで、検索要求にマッチする記事とそれに関連する記事群を得ることができる。

3.2.2 提示する特徴語の選択

まず、一覧表示する記事すべてに共通する語 (複合語単位) を主要トピックとして取り出し、各記事の特徴語の候補から外す。主要トピックは記事一覧の上に表示する。

類似度順に一覧表示する各記事の表示内容は、見出し、URL、および特徴語である単語のリストである。なお、これらの単語はチェックボックスにより選択ができるようになっており、次の検索の検索語として用いることができる。

さて、各記事の「特徴語」とは、差分を複合語で取ることで得られる複合語の集合 P_i である。差分を複合語で取るとは次のように定義される。記事を上位から順に $A_1, A_2, \cdots$ とする。i 位の記事 A_i に複合語 $W_c(= W_{s_1}, W_{s_2}, \cdots, W_{s_k}(W_{s_i}(i = 1 \cdots k)$ は単純語)) が出現した場合、それよりも順位の低い記事 $A_j(j > i)$ では W_c に一致する複合語の表示はしない。しかし、W_c の構成要素 $W_{s_1}, W_{s_2}, \cdots, W_{s_k}$ を含む、W_c とは異なる複合語は表示する。この複合語を構成する単純語の記事内での TFIDF を計算し、その値を平均したものを複合語の重みとする。この重みの大きい順に特徴語として提示する。

4 「いもなび」の評価

「いもなび」は、関連トピックを俯瞰したい場合、意外性のある関連性を見つけたい場合、単に興味のあるページをさまよいたい場合など、さまざまな場面での利用が考えられるが、それらを実現する基盤となっているのは、メインのトピックから周辺のトピックへとユーザを連続的にナビゲートするトピックドリフトの機能である。そこで、「いもなび」のトピックドリフトの性質を明らかにするための実験を行なった。

システムがうまく周辺のトピックへとユーザをナビゲートできれば、検索により適度に離れたトピックの記事が得られているはずである。そこでトピックドリフトの指標として、検索結果で 1 位の記事と、その次の検索結果で 1 位の記事との類似度を用いることにした。

各種条件による類似度の違いを評価 A(実験 1〜実験 5) で、記事の類似度と記事内容の関連度との関係を評価 B で明らかにする。

4.1 評価 A: 各種条件による類似度の違い

実験の手順

1. あらかじめ定めた検索語を入力し検索
2. 検索結果の類似度 1 位の記事を得る
3. 検索結果で提示された特徴語をランダムに選択し、再び検索
4. 検索結果の類似度 1 位の記事を得る
5. 今得られた類似度 1 位の記事と 1 回前に得られた類似度 1 位の記事間での類似度を求める
6. 3〜5 を 10 回繰り返す

この手順により、1 単語につき 10 回分 (10 個) のデータを得る。この実験を 20 単語、30 単語、50 単語で行い、それぞれ 200 個、300 個、500 個のデータを得る。なお、実験には 1 週間分の記事を使い、最初に入力する単語は主要新聞社 5 社の記事の見出しから無作為に抽出した。検索結果の画面に表示される記事の数は最大 5、1 記事につき表示する特徴語の数も最大 5 とした。3 でランダムに選択する特徴語は全体の 10%(表示する最大 25 単語の内の 2〜3 単語) である。

実験の条件

実験 1：**デフォルト条件による実験**
システムのデフォルトの条件である「検索に用いる記事データは主要 5 社の新聞社記事データ、2 段階の並べ替えを行う、記事間の差分を複合語で取る」という条件で実験を行う。他の実験は、この実験 1 の結果との比較で議論する。

実験 2：**検索に用いる記事データ群の差による実験**
少数の新聞社の記事データと多数の新聞社の記事データでは検索結果にどのよう差が生じるのかを検証する。検索には Google ニュース [2] から得た記事データを用いる。Google ニュースは同一内容の記事がまとめられており、また、610 以上のサイトの記事を持つため、新聞社 5 社の記事と比べひとつのトピックに対して多くの関連記事を得られる。

実験 3：**記事間の差分を取らない実験**
検索後に各記事 $A_i(i = 1,2,\cdots)$ の特徴語を差分を取らずに提示する。すわなち 2 節の記法において $W_i(i = 1,2,\cdots)$ を A_i に対応して表示した場合での実験を行いデータを得る。差分を取らないことにより、それぞれの記事で TFIDF の値が高い単語がそのまま特徴語となる。

実験 4：**2 段階の並べ替えを行わない実験**
検索後に提示する記事を 3.2.1 節で述べた 2 段階の並べ替えを行わずに提示した場合での実験を行う。2 段階の並べ替えを行わないと、検索要求との類似度順で記事が並ぶため、1 位の記事の関連記事が集まりにくくなる。

実験 5：**差分を単純語で取る実験**
特徴語を複合語でなく単純語で差分を取り提示した場合

での実験を行う。差分を単純語で取るとは次のように定義される。記事が上位から順に $A_1, A_2, \cdots$ と並ぶとする。記事 A_i に語 $W_c(= W_{s_1}, W_{s_2}, \cdots, W_{s_m}$ ただし $W_{s_i}(i = 1 \cdots m)$ は単純語) が出現したとき、$W_{s_1}, W_{s_2}, \cdots, W_{s_m}$ のすべてが上位記事 $A_j(j < i)$ で出現している場合のみ排除される。上位記事に W_c 全体として出現していない場合でも排除されることがある。

以上の実験条件をまとめると表 1 になる。

表 1: 実験条件

	データベース	差分の有無	並べ替えの有無	差分の単位
実験 1	新聞社 5 社	差分あり	並べ替えあり	複合語
実験 2	Google ニュース	差分あり	並べ替えあり	複合語
実験 3	新聞社 5 社	差分なし	並べ替えあり	複合語
実験 4	新聞社 5 社	差分あり	並べ替えなし	複合語
実験 5	新聞社 5 社	差分あり	並べ替えあり	単純語

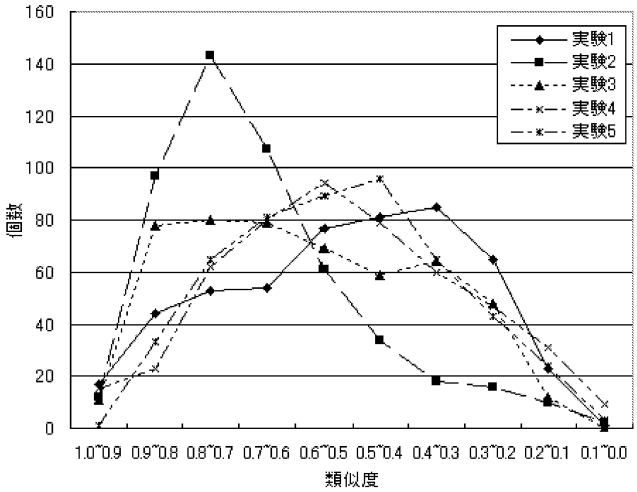

図 6: 検索前後の 1 位の記事の類似度 (データ数 500)

実験結果

実験 1〜5 の結果を図 6 に示す。なお、データ数 200、300、500 それぞれの結果を比べた結果、どれも傾向が同じであった。このことから、これ以上データを増やしたとしても結果に大きな差はでないと判断できる。ここではデータ数 500 の結果を示し議論する。

実験 1 と実験 2 の比較より、類似記事の量がドリフト量に影響を及ぼすことがわかる。多数の新聞社の記事データを用いると同一内容の記事が多くなり、次の検索でも同一内容の記事が 1 位となる確率が高くなっていると考えられる。

実験 1 と実験 3 の比較より、差分を取ることでより遠くへドリフトすることがわかる。差分を取らない場合は各記事における TFIDF 値の高い語に重なりが見られ、またそれらが元のトピックに近いことが多いため、トピックが移り変わりにくくなっていると考えられる。

実験 1 の結果と実験 2〜5 の結果に差があるかどうかを t 検定によって調べたところ、棄却率 95%において、実験 1 と実験 4、および実験 1 と実験 5 では結果に差があるとは言えないと判定された。記事の並び替えの有無と、差分を取る際の複合語/単純語の違いは類似度に影響しないという結果である。ただし、並び替えの有無に関しては人間が観察していると表示される記事になんらかの差があるため、類似度では判定できない違いがあるものと考えられる。

4.2 評価 B: 類似度と記事内容の関連性との関係

評価 A の実験 1 と同じ条件で実験を行ない、検索前後の記事を人間が読んで、その関連度を「同一内容の記事」「関連があるが異なる内容の記事」「関連の全くない記事」の 3 段階で判断する実験を行なった。データ数は 100 である。

実験結果は図 7 となり、類似度 0.5 以下はほぼ無関係の記事ということがわかる。評価 A の結果と照らし合わせると、元のトピックとは関連のないトピックへとドリフトしていることも多いことがわかる。

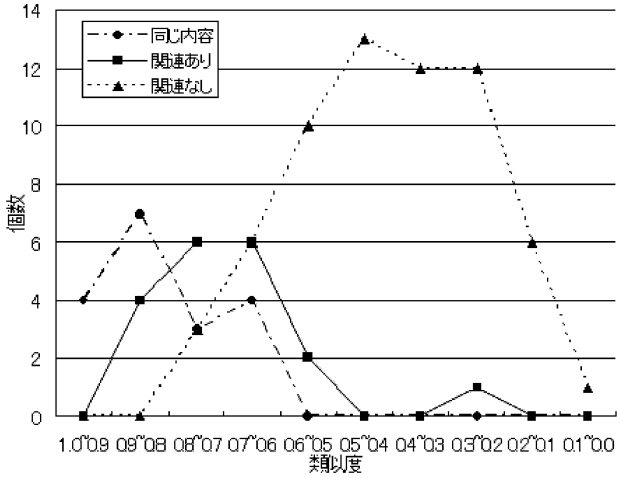

図 7: 記事の類似度と人間が判断した関連性

5 おわりに

本研究では関連トピックをいもづる式に得ることができる検索エンジンを実装し、そのトピックドリフトの性質について評価実験を行なった。その結果から、本研究のシステムは記事の差分の単語をナビゲーションに用いることにより、より類似度の小さい記事へとナビゲートし、関連性のない記事へドリフトすることもあることがわかった。類似度の値、関連性の有無と、ユーザの満足度との関係を分析することが今後の課題として挙げられる。

参考文献

[1] 奥村学, 難波英嗣 (2002)「テキスト自動要約に関する最近の話題」『自然言語処理』9(4), pp.97-116.

[2] Google. Google ニュース 日本版. URL: http://news.google.co.jp/.

本研究は，社会技術研究システム ミッション・プログラム「安全性に係わる社会問題解決のための知識体系の構築」(2001〜2002 年度は日本原子力研究所の事業，2003 年度からは科学技術振興事業団の事業) の研究として行われた.