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」の評価

西端 紳一郎　　山田 剛一　　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
{nisihata@cdl.,yamada@,masuda@}im.dendai.ac.jp, nakagawa@dl.itc.u-tokyo.ac.jp

1　はじめに

インターネット上には，複数の新聞記事サイトなどから日々大量の記事が無料で配信されている．その中から，「あるトピックの記事を読み，さらに関連するトピックについても知りたい」というユーザの要求に応えることが望ましい．

そこで我々は検索エンジン「いもなび」[1] を開発した．このシステムでは，複数の新聞記事サイトを横断的に検索し，さらに関連するトピックを知りたいという欲求に応えるよう，検索結果の表示において，記事間の差異を各記事のトピックを示す特徴語として単語で提示する仕組みを組み込んでいる．そして，この単語を用いて再検索することができる．これによりユーザは関連する情報を容易に得ることができる．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」に情報が得られる．通常の検索エンジンが提供する一回限りの検索に対し，この連続的な検索こそが，本システム「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．

本研究では，このナビゲーション機能に重点をおいてシステムの評価を行った．

2　システムの概要

本システムは，複数の新聞記事サイトを横断検索すると共に，ユーザの検索ナビゲートを行う．

複数の新聞記事サイトを横断検索することで，内容が類似した記事群を得ることができる．この記事群において，共通している情報や，記事間の差異の情報を求め，ユーザに提示する．

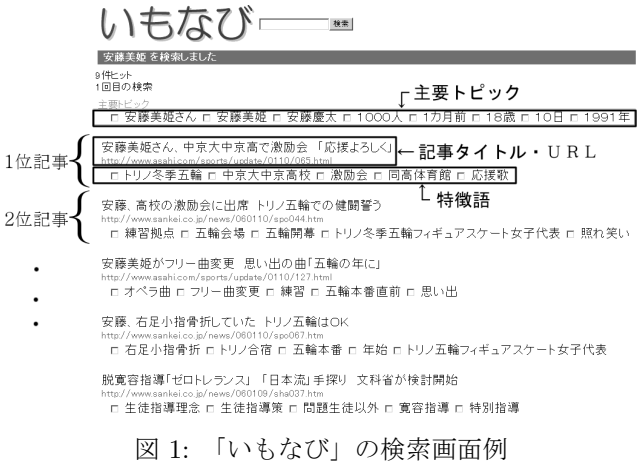


図 1: 「いもなび」の検索画面例

図 1 は,「いもなび」による新聞記事の検索結果画面である．検索結果では，記事のタイトルや URL と共に，検索結果の記事に共通して含まれる単語を「主要トピック」，記事間の差分の単語を「特徴語」としてユーザに提示している．ユーザは，検索結果で得られた記事を見ると共に,「主要トピック」や「特徴語」の中から興味のある単語を選択して，さらに検索を行うことができる．図 2 が，新たなトピックへドリフトした例である．図 2-(a) で検索した結果で得られた「主要トピック」や「特徴語」を選択し，図 2-(b) のように新たなトピックへドリフトする．

「主要トピック」や「特徴語」の中から，ユーザ自身が興味のある単語を選択して検索を繰り返すことで，興味のある新たなトピックへ検索ナビゲートされるのである．

本システムの構造の詳細については，以前に述べている [1] ので省略する．本システムでは，主要新聞社 5 社の記事を収集し，検索の対象としている．

2.1　記事の検索アルゴリズム

あらかじめ収集している記事の中から，検索要求を満たす記事群を抽出する．その中から，検索要求に最も近い記事と，それに関連する記事を求めるために 2 段階の類似度計算を行う．

まず，検索要求の単語群と，検索要求を満たしている各記事との間で類似度計算を行い，類似度の値が最も高い記事を算出する．次に，検索要求に最も近い記事と，検索要求を満たす残りの記事群の間で再び類似度計算を行う．これにより，検索要求に最も近い記事と，その記事に関連する記事群を求めることができる．

この 2 段階の類似度計算には，ベクトル空間法を用いている．ベクトルの次元は，検索要求の単語群および各記事に含まれる単語群の種類数である．また，検索要求のベクトルの要素は検索に用いた全記事に対するその単語の IDF 値を用い，各記事のベクトルの要素は各記事におけるそれぞれの単語の TF 値を用いている．

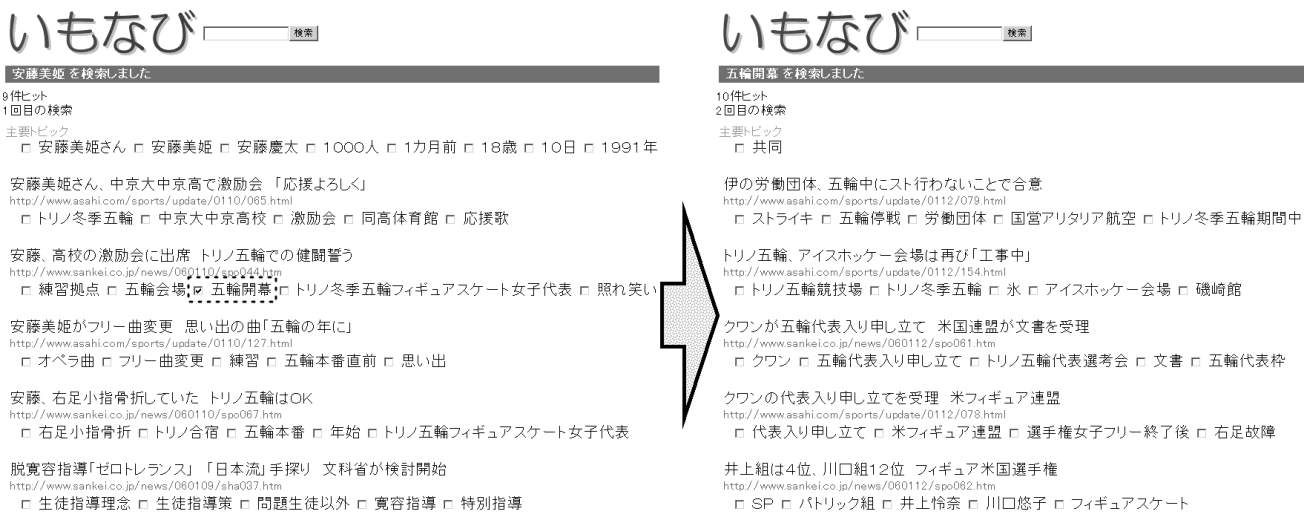
さらに，ユーザに提示する記事に共通して含まれる単語を「主要トピック」として扱う．また，各記事に含まれる単語の TFIDF 値を算出し，値が大きいものを各記事の「特徴語」として扱う．ただし，この特徴語は，その記事より上位の記事の特徴語に含まれているものは排除する．この「主要トピック」および「特徴語」は，記事中で複合語として現れている場合は，その複合語のまま扱う．

以上によりユーザに提示する情報を算出したら，検索結果として図 1 のような画面でユーザに提示する．検索結果は上位 5 位までの記事，各記事の特徴語は上位 5 個まで提示する．

ユーザは「主要トピック」や「特徴語」の中から，任意に単語を選択してさらに検索を行うことができる．システムは，この選択された単語群を新たな検索要求として記事の検索を行う．

3　評価実験

「いもなび」は，関連するトピックを俯瞰したい場合や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．これらを実現する基盤となっているも



(a) ドリフト前　　(b) ドリフト後

図 2: トピックのドリフト例

のは，メインのトピックから周辺のトピックへと，ユーザの検索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．また，この機能はユーザ自身が興味の趣くままに情報を検索することを目指している．つまり，本システムでは，

- 関連性のある情報が得られたか
- 興味のある情報が得られたか

の 2 つが重要なポイントとなる．そこで今回は上記 2 つのポイントに焦点を絞って実験を行った．

3.1 トピックのドリフト量

この実験では，本システムのナビゲーション機能により，どの程度トピックがドリフトした記事が得られるのかを調査した．ドリフトの程度は，検索結果 1 位の記事とその次の検索結果 1 の記事との間の類似度の値を用いることにした．つまり，類似度が高ければあまりトピックがドリフトしておらず，類似度が低ければ，大きくドリフトしていることになる．

また，本システムのナビゲーションには，以下の 2 つが大きく影響を与えると考えられる．

- ユーザが選択する単語群の変遷
- システムが提示する記事の変遷

そこで，これらの要因も考慮するために，以下のような実験を行った．

実験の手順

1. 決められた検索語を入力し検索
2. 検索結果の 1 位の記事を得る
3. 検索結果で提示された特徴語を選択しふたたび検索
4. 検索結果の 1 位の記事を得る
5. 今得られた 1 位の記事と 1 回前に得られた 1 位の記事間でのコサイン類似度を求める
6. 3〜5 を 5 回繰り返す

この手順により，1 つの検索語に対し 5 個のデータを得る．ユーザが選択する単語群の変遷に伴うナビゲーションの影響を見るために，手順 3 において，被験者が任意に特徴語を選択する場合と，機械がランダムに特徴語を選択する場合とで比較を行う．選択する特徴語の数は 2〜3 個とした．

被験者は，理工系大学に通う学生で日頃から Web 検索には慣れている人を用いた．

検索対象は，主要新聞社 5 社から 1 週間で取得した記事とした．記事数は約 3000 記事である．

また，システムが提示する記事の変遷に伴うナビゲーションの影響を見るために，検索対象の記事を，Google ニュース [2] より取得した記事データを用い（以下，Google 版），同様の手順で実験し比較する．検索の対象は，1 週間で取得した記事とし，記事数は約 7000 記事である．Google ニュースは記事が関連ごとにまとめられており，また，610 以上のサイトから記事が集められているため，主要新聞社 5 社だけ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ることができる．

実験結果

図 3 及び図 4 より，実験の手順 3 において特徴語をランダムにチェックして検索した結果よりも，特徴語を被験者が選択して検索する方が，類似度の低い記事が多く検索されている．また，図 3 と図 4 を比較すると，検索対象の記事が多い図 4 の方が，類似度の高い記事が多く検索される傾向にあることがわかる．これは，Google 版では，主要 5 社版よりも多く数の記事が検索の対象となるため類似度の高い記事も多くなるためであると考えられる．

3.2 類似度と関連性の関係

3.1 節では，「いもなび」のナビゲーション機能によるトピックドリフトの程度を，類似度を指標に実験を行った．

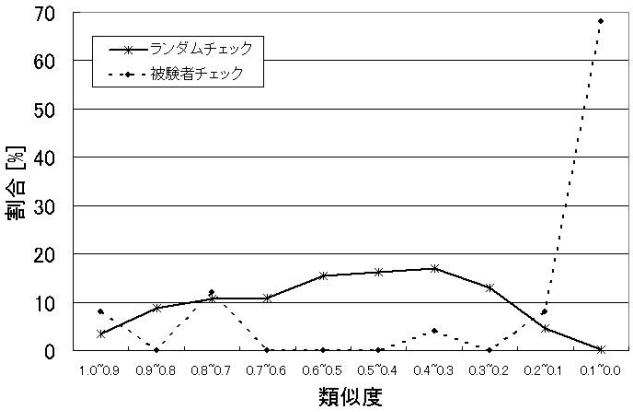


図 3: 選択する単語の変遷が与えるドリフトへの影響（主要 5 社版）

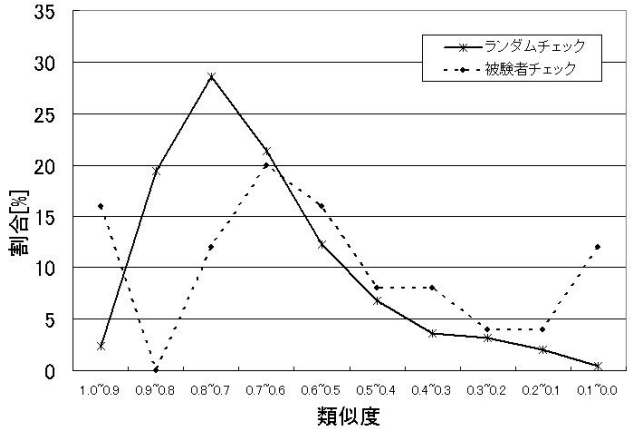


図 4: 選択する単語の変遷が与えるドリフトへの影響（Google 版）

そこで，記事間の類似度の値と，人間が捉えるトピックの関連性について調査した．

実験の手順

1. 決められた検索語を入力し検索
2. 検索結果の 1 位の記事を得る
3. 検索結果で提示された特徴語を被験者が任意に 2〜3 個選択しふたたび検索
4. 検索結果の 1 位の記事を得る
5. 今得られた上位 5 位までの各記事と 1 回前に得られた 1 位の記事間でのコサイン類似度を求める
6. 今得られた上位 5 位までの各記事と，1 回前に得られた 1 位の記事との間でそれぞれ話題に関連性があるかを被験者が評価する
7. 3〜6 を 5 回繰り返す

この手順の実験を被験者に行ってもらう．手順 6 では，

- メインの話題が同じである
- メインの話題に関連性がある
- その他

の 3 段階の評価とした．この実験でも，システムが提示する記事の変遷に伴う影響を見るために，検索対象となる記事を主要 5 社によるものと Google ニュース [2] によるもので実験を行った．被験者，および検索対象の記事は 3.1 節の実験と同様である．

本実験，および 3.3 節の実験は，被験者が評価を入力できる機能を「いもなび」に組み込んで実験を行った．

実験結果

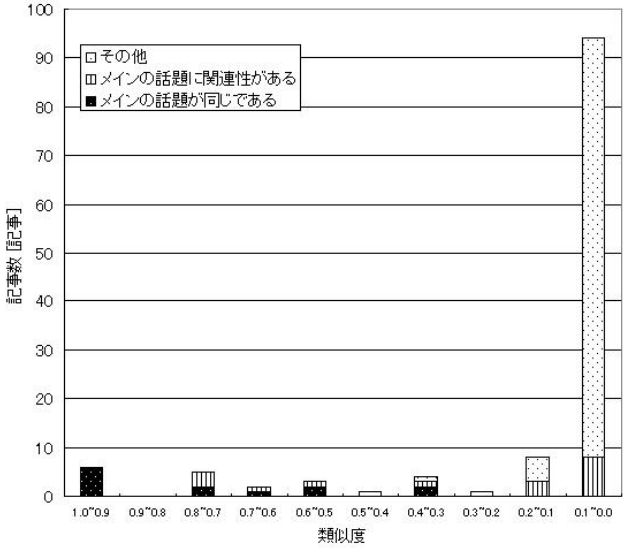


図 5: 類似度と関連性の関係の主観的評価（主要 5 社版）

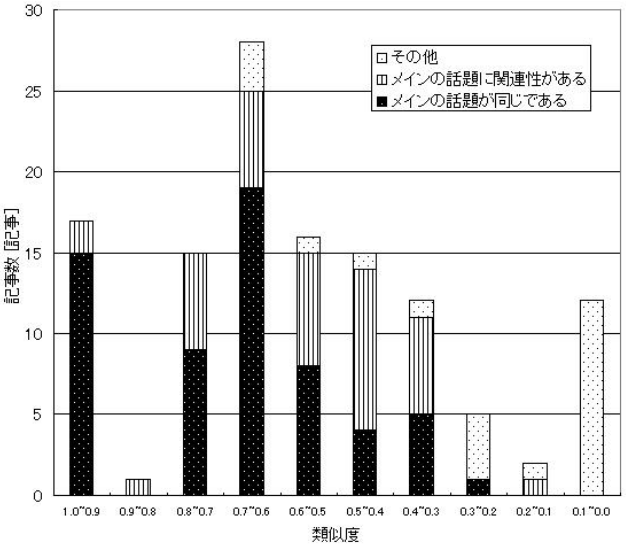


図 6: 類似度と関連性の主観的評価（Google 版）

図 5 および図 6 より，両グラフともに記事間の類似度が高いものほど被験者は「話題が同じである」や「話題に関連性がある」と評価し，類似度が低いものは被験者は「その他」との評価が多くなっている．このことから，機械的に求められる「記事間の類似度」と，「人間が評価する記事間の話題の関連性」には，一定の関係性があるといえる．

3.3 類似度と記事に対する興味の度合いの関係

「いもなび」のナビゲーション機能では，ユーザが興味の趣くままに検索できることが重要である．そこで，本システムによる検索によって，本当にユーザが興味のある記事を検索できているのかを調査した．

実験の手順

3.2 節の実験の手順 6 において，得られた上位 5 位までの各記事に対し，被験者がその記事に対する興味の度合いを評価する．興味の度合いは，

- 興味深い : 今すぐに記事の本文を読みたいと思う記事
- 興味あり : 時間があれば記事の本文を読みたいと思う記事
- 興味なし : 記事の本文を読みたいとは思わない記事

の 3 段階とした．被験者，および検索対象の記事は 3.1 節の実験と同様である．

実験結果

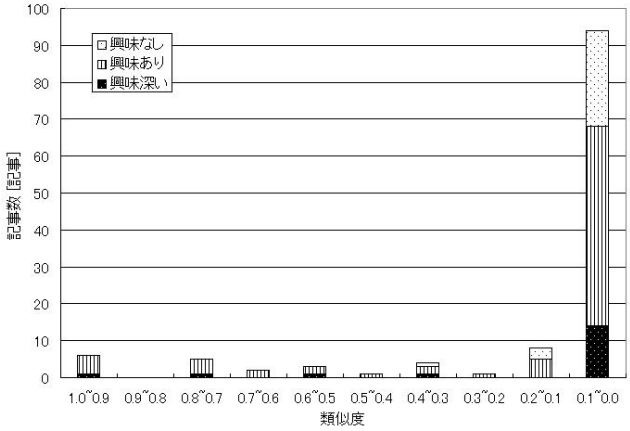


図 7: 類似度と記事に対する興味の度合いの主観的評価（主要 5 社版）

図 7 および 8 より，類似度 0.4 以下において，被験者の記事に対する興味の度合いが「興味深い」または「興味あり」と評価されている記事が，主要 5 社版では 72.0%，Google 版では 87.1%であった．一方，類似度が 0.6 以上の場合は，主要 5 社版ではデータ数が極端に少ないが，Google 版では 32.8%となった．つまり，被験者は類似度が低い記事に対してより興味を抱きやすいことがわかる．

4 考察

本システムは，3.1 節の実験結果より検索対象となる記事数がある程度多い場合は，類似度の高い記事を検索することができるということがいえる．しかし，3.3 節の結果より，被験者は類似度の高い記事よりも，類似度の低い記事に対して興味を抱く傾向があることがわかった．このことから，本システムの利用者は，あまり類似した記事ばかりが検索されるよりも，検索される記事のトピックが大きくドリフトすることを期待していると考えられる．また，3.2 節の結果より，記事間の類似度と被験者が評価した記事間の話題の関連性には，ある程度の関係性が認められた．このことから，本システムが類似記事を求める際に，類似度計算の結果を元に，類似記事を提示するという手法が妥当であったと考えられる．

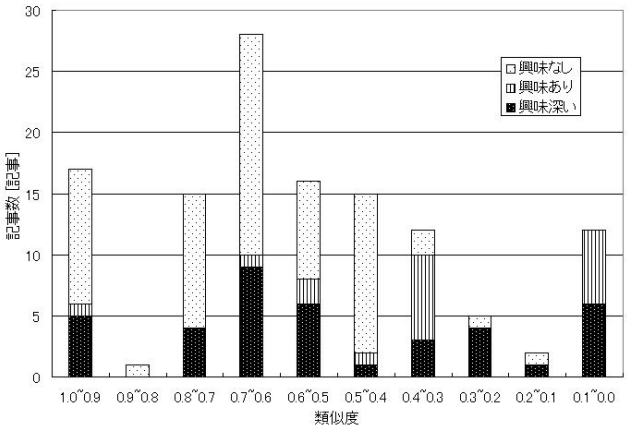


図 8: 類似度と記事に対する興味の度合いの主観的評価（Google 版）

5 おわりに

本研究ではトピックドリフトを支援する新聞記事検索システムを構築し，システムの評価を行った．

本システムは，類似度の高い関連する記事をいもづる式に検索することができるということが評価実験の結果より言える．一方で，ユーザが本システムを利用した場合，関連性の高い記事よりも，関連性の低い記事を検索する傾向が高く，また，そのような記事に対して興味を抱く場合が多いことがわかった．

しかし，その理由については，今回の実験では明らかになっていない．我々は，その理由の一つに，本システムのユーザは「意外な記事」が検索されることを期待して検索を行い，そして，検索された「意外な記事」に対して興味を抱くためではないかと考えている．今後は，記事に対する興味と意外性の関係について調査を行いたい．

参考文献

[1] 山田剛一, 大熊耕平, 増田英孝, 中川裕志 (2005)『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」』言語処理学会第 11 回年次大会 B5-9

[2] Google ニュース. 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．