DEWS2005 4B-i11

Web検索結果のクラスタリングと観点抽出に基づく閲覧インタフェース

八村　太輔†　　湯本　高行††　　赤星　祐平††　　小山　聡††　　田中　克己††

† 京都大学工学部情報学科
〒 606-8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科 社会情報学専攻
〒 606-8501 京都府京都市左京区吉田本町
E-mail: †{hatimura,yumoto,akahoshi,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　近年，Web 上に存在する情報量が飛躍的に増加するにつれ，検索エンジンはこの膨大な情報の中からユーザの意図に合うものを，効率的に絞り込むために必要不可欠なツールとなった．しかしある検索者が，自分の関心を持つトピックの動向や変遷を Web から取得しようとするとき，検索エンジンは必ずしも効果的な手段とはいえない．なぜなら，一般的に検索エンジンより取得される検索結果は URL や文章の断片を羅列したものでしかなく，また各々のランキング手法によって人気のあるページが常に上位に位置づけされやすく，トピックの全体的な動向はこれらの中に埋没してしまう傾向にあるためである．そこで本論文では，検索エンジンより取得される検索結果の時間変化を考慮する，検索結果ページ集合のバースト度を定義する．バースト度は，ある記事が前回の検索実行時と比べて，どれだけ主要度が高いかを判断する指標となる．またこのバースト度に基づき，トピックの動向変遷の効率的な閲覧を支援するインタフェースとして，新聞メタファを実装する．
キーワード　情報検索，ユーザインタフェース，Web 利用技術

A Browsing Interface for Web Search Results by Data Clustering and Aspect Discovery

Taisuke HACHIMURA†, Takayuki YUMOTO††, Yuhei AKAHOSHI††, Satoshi OYAMA††, and Katsumi TANAKA††

† School of Informatics, Faculty of Engineering, Kyoto University
Yoshida-hommachi, Sakyo-ku, Kyoto, 606-8501 Japan
†† Graduate School of Informatics, Kyoto University
Yoshida-hommachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †{hatimura,yumoto,akahoshi,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　Recently, a search engine becomes an effective solution to retrieve informations from the Web, with the rapid spread of Internet. However, it is not a suitable way to aquire the changes or trends of a certain topic using the search engine. It is because thier search result are nothing but a enumeration of URL and some fragments of sentences, and moreover, the popular pages always rank among the top of search result. The overall changes or trends of a topic are hard to grasp on that account. Therefore, we propose a method for detecting the burstly occuring topic by examining the fluctuation of search results. Then we propose a new interface for browsing the burstly occuring topics. This interface implement Newspaper metaphor by data clustering and aspect discovery.
Key words　Information Retrieval, User Interface, Technology using the Web

1. はじめに

近年のインターネット環境の普及に伴い，Web 上に存在する情報量は飛躍的に増加し，ユーザが一度に大量の情報に接する機会が増えてきた．このような環境下において，ユーザが必要としている情報だけを検出するのは非常に困難な作業であるといえる．

Google [6] などに代表される検索エンジンはこの膨大な情報

の中から，ユーザの意図に合う情報を効率的に絞り込むため，必要不可欠なツールである．しかし Web 空間に存在するあまりに多くの情報量のために，検索エンジンを使用して情報を絞り込んでもなお，その検索結果が数百件に及ぶことは頻繁に起こりうる．これらの検索エンジンから得られる検索結果は，ユーザの与えたキーワードを含むページの URL と文章の一部分を抜き出して羅列するのみで，そこに本当に欲しい情報が載っているかどうか，実際にそのページをひとつひとつ開いて閲覧しなければ判断できない．

さらに Google に代表される多くの検索エンジンでは，PageRank などの Web ページ間のリンク関係に基づいた順位付けを行っているが，こうした順位付けは潜在的にもともと有名であるページをより上位に配置してしまいがちである [5]．結果的に，ユーザが接することのできる情報は，検索結果のうち上位に位置する，一部のものだけになってしまうという可能性を免れない．

そこで本論文では，検索エンジンより取得される検索結果の時間変化を考慮して，文書集合の主要度判定を行う手法を提案する．またこの主要度判定に基づき，トピックの変遷や動向の効率的閲覧を支援するインタフェースとして，新聞メタファを実装する．新聞メタファを採用する利点として，以下のような点が挙げられる．

新聞は普通，互いに漠然と結びついた多様な記事の中から，選択的に読者が記事を閲覧できるよう構成されている．そしてその一面には複数の記事の概要が，より詳細な記事への参照とともに掲載される．また記事の配置は記事同士の相対的な重要性に依存し，重要度の高い記事ほどトップに掲載され，より大きく扱われる傾向にある．このように記事をレイアウトすることで，読者に対し潜在的に有用である情報を喚起することができる．

また新聞は，定期的な刊行物であるという特徴も有している．本研究で提案するシステムは，検索エンジンに対して定期的にクエリを投げることで時系列順に検索結果を取得し，その時間的変化に注目しコンテンツの重要度判定を行う．この重要度に基づいてコンテンツを提示することで，ユーザは記事の動向や変遷を把握しやすくなる．

以降，第 2. 章では提案手法の関連研究を挙げ，第 3. 章で新聞メタファの概要と全体の概観を述べる．第 4. 章で検索結果から記事の抽出手法について，第 5. 章で検索結果ページ集合のバースト度について，第 6. 章でプロトタイプとそれに基づく本研究の考察，第 7. 章で結論を述べる．

2. 関 連 研 究

2.1 Burst 検出

Kleinberg [1] は，ある時間間隔をおいて複数の文書が到着するような文書ストリームについて，特定のトピックに関する文書の到着頻度に着目し，その活性度を発見する手法を提案した．

あるトピックが活性状態にある時，当該トピックに関する文書の到着時間間隔は確率的に短くなる．この状態を Burst 状態と定義する．Kleinberg は，文書の到着間隔はポワソン分布に

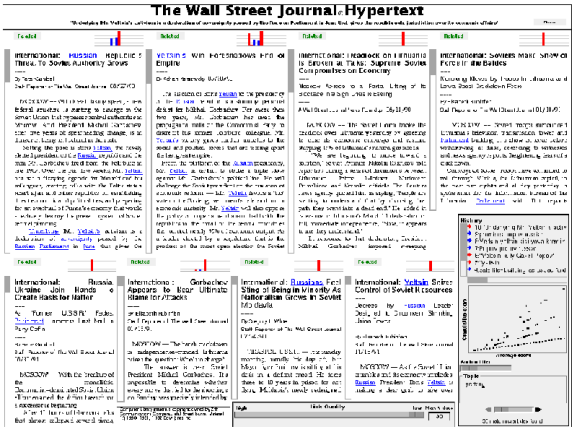

図 1　VOIR による新聞メタファのプロトタイプ

従うものであり，その到着時間の平均間隔は隠れマルコフモデルにより確率的に出力されると仮定している．この隠れマルコフモデルにおいて，平常状態とバースト状態の間を遷移する際のコストの総和を最小化する状態遷移列を求めることにより，あるトピックの時間経過に伴う活性度の変化を調べる．

Kleinberg による Burst 検出は，文書ストリームを対象にする場合に効果を発揮するが，Web アーカイブなどを持たない一般ユーザがこの手法により文書ストリームを監視するのは困難であると思われる．本研究では，検索エンジンに対して定期的にクエリを投げることで，Web 検索結果集合のバースト度を定義することを考える．

2.2 TDT

TDT(Topic Detection and Tracking) [2] は，時系列順に配送される各種ニュース記事の分類やスレッドの作成，トピックの検出などを目的とする，評価型の会議である。TDT では問題を以下の 5 つに分割し，研究を進めている．

- **Story Segmentation** トピックの切れ目を検出
- **Topic Tracking** 指定したトピックと同一トピックをトラッキング
- **Topic Detection** 同じトピックを持つ文書を検出
- **First Story Detection** 新規トピックの出現を検出
- **Link Detection** 二つのトピックが等しいかどうか検査

TDT も Kleinberg による手法と同様にストリームを対象とするもので，特にオンラインの情報から実時間によるトピックの検出を扱うものが多い．本研究では Web 検索結果からのトピック検出を目的とするため文書ストリームは使用しない．また対象は必ずしもニュース記事に限らず，ユーザが与えた任意のクエリから取得される検索結果を元に新聞を生成する．これらの点において，本研究による手法は TDT と異なっている．

2.3 VOIR

VOIR(Visualization of Information Retrieval) [3] とは，Gene Golovchinsky によって提案された，動的ハイパーテキストを用いた新聞インタフェースの手法である．新聞とは互いに漠然と結びついた多様な記事から構成され，普通その一面には記事の概略が，より詳細な記事の載った紙面への参照とともに掲載される．VOIR はこうした記事間における関係をハイパー

リンクに見立て，自動的に文書間にハイパーリンクを作成することで，ユーザのナビゲーションを支援する．

ユーザは文章を選択したり，クエリを打ち込んだり，ハイパーリンクを選択することで，閲覧の意図を VOIR に伝える．VOIR はこうしたユーザの閲覧動作から，ユーザの関心が高いと思われる単語をリンクアンカーとして抽出し，全文検索エンジンを使って他の文書へのハイパーリンクを自動的に作成する．

本論文では時系列順に得られる Web 検索結果から，記事間の相対的な重要性を判断し，新聞インタフェースにより検索結果を表示することを主題としている．一方 VOIR の目標は，文書間における動的なハイパーリンクの作成によって，ユーザの閲覧を支援することであり，この点で本論文とは異なっている．

3. 新聞メタファ

本研究の目標は，検索エンジンから得られる検索結果を時系列順に解析することで，その主要性を判断することであり，そのためのインタフェースとして新聞メタファを提案する．これは新聞の網羅性・定期刊行性という二つの性質を利用することにより，トピックの動向や変遷の把握を支援するものである．以下に，新聞の網羅性と定期刊行性について説明し，さらに提案システムの概要を示す．

3.1 新聞の性質

3.1.1 網 羅 性

一般的に新聞は多様な記事から構成され，選択的に読者が閲覧しその内容や動向を把握しやすいように配置されている．まず第一面には，発行された時点において話題性の高い複数の記事の概要が網羅的に配置される．また一面記事は，より詳細な記事への参照とともに掲載されるため，ある記事についてさらに多くの情報を得たいと思うユーザはこの参照を辿って，他の面へ移動すればよい．

記事の配置手法は，記事同士の相対的な主要性に依存している．つまり主要度の高い記事ほどトップに掲載され，より大きな紙面を割り当てられる．このようなレイアウトは，ユーザに対して潜在的に情報の有用性を喚起することができる．

特にインターネットで配信されるニュースと紙の新聞を比較すると，ユーザが自ら欲しい情報を取捨選択できるかどうか，という点で大きく異なると思われる．インターネットで配信されているニュースは，ユーザが読みたい情報のみを効率的に選び出して閲覧できる特徴を持つが，かえってそのために知らなかった情報の重要性に触れる機会が減ってしまうのではないかと考えられる．一方で新聞の第一面には，様々なトップ記事が網羅的に紙面上に配置されているため，こうした発見の機会を見逃さずに済むであろう．

3.1.2 定期刊行性

新聞は定期的に発行され，ユーザの元へ届けられる．この時，新聞は発行された時点において随時関心度の高いトピックを記事として扱うので，常に新しく重要な情報を取得することができる．このためユーザは刊行された記事を時系列順に閲覧することで，トピックの動向や変遷を追うことが可能になる．

本研究は，定期的に取得される検索結果集合を解析することで文書集合の主要度を判定するため，こうして得られた記事の動向や変遷を把握する手法として，この新聞メタファは有効であると思われる．

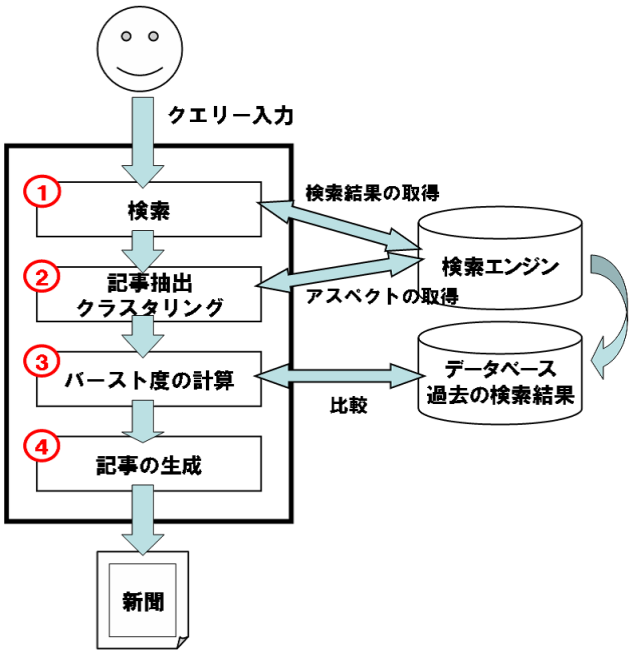

図 2 システムの流れ図

3.2 シ ナ リ オ

今,「スマトラ沖地震」と言う事項に対して関心を寄せている Web 検索者を想定する．当初スマトラ沖地震に関するトピックは,「スマトラ沖で大地震が起こった」という第一報から始まった．ところが時間の経過とともに被害状況が明らかになるにつれ,「日本が 5 億ドルの資金援助」という経済面や,「被災地で伝染病蔓延の恐れ」といった保健面での動向を伝えるトピックなど，複数の観点を持つトピックへと変化していった．

スマトラ沖地震について関心を持つ検索者は，検索エンジンに「スマトラ沖地震」というキーワードを入力し，このクエリにヒットする検索結果を得ることで，事件の概要と動向を知ろうとするであろう．しかしどのタイミングで検索しても，得られる検索結果で上位に位置するものの内，そのほとんどが被災地への救援を呼びかける記事である．つまりトピックの動向や変遷を検索によって把握するのは，従来の検索エンジンでは不足であるといえる．

新聞メタファを用いる本手法は，新聞の網羅性や定期刊行性といった特徴を利用することで，この問題を解決する．ただしここで定義する新聞メタファは，必ずしも事件や事故といったニュースのみを新聞形式で表現するものではない．ユーザが興味を持つ任意のキーワードをシステムに与え，得られる検索結果から記事を生成して提示するため，例えば「花粉新聞」や「携帯電話新聞」などの新聞を生成することも可能である．

本研究で提案するシステムは，以下の流れによって新聞を構成していく．

3.2.1 記事の検索

ユーザは自分が関心を持つトピックについて，そのキーワー

ドをあらかじめシステムに入力しておく．システムは与えられたキーワードをクエリとし，定期的に検索エンジンから検索結果を得る．これらの検索結果は時系列順に取得されるため，その時間変化を調べることでトピックの動向や変遷を得ることができる．

3.2.2 記事の抽出

得られた Web 検索結果の文書群の中には，同一のトピックについて言及している文書が複数存在していると考えられる．これらを検索結果のリンク集からひとつひとつ閲覧していくのは，非常に効率が悪い．そこで，同一のトピックを持つと思われる文書を，その類似性によってクラスタリングし，ユーザの要求する記事数分だけクラスタを生成する．得られる各クラスタをひとつの記事と定義し，それらを紙面上に配置する．

さらに各クラスタ内に存在する各文書について，そのリンク元ページを抽出し，これらのリンク元ページから対象ページに対する観点を取得する．一般的に，あるページに対するリンク元ページには，当該ページに対する社会的評価が記述されているものと考えられる．こうした記述をリンク元ページから抜き出し，その中から大きな特徴量を持つキーワード集合を対象ページの観点と定義する．ここで得られた観点が，紙面上の ”経済面 ”や ”スポーツ面 ”といった，記事の属性に対応するものであり，この観点に基づき各ページは分類される．

3.2.3 バースト度の判定

前項のクラスタリングによって得られる各記事の主要度を示す指標として，検索結果ページ集合のバースト度を計算する．このバースト度とは，現時点での検索結果とそれまでに得られた検索結果の時間的変化を評価するものである．

まず，検索結果のクラスタリングによって得られた各記事について，過去の記事集合との対応を求める．これは現時点における記事の特徴ベクトル（当該記事に対応するクラスタ内の全文書の特徴ベクトルの重心）との類似度が最も高い，前回の検索結果中の記事を選択することで得られる．次にこの対応する記事同士で，その記事を構成するクラスタの文書数や文書の被リンク数などの時間的変化を調べることで，注目するクラスタのバースト度が計算される．

バースト度は，第 5. 章で詳しく定義する．

3.2.4 記事の生成

紙面を生成する際，全記事をそのバースト度によってソートし，上位からユーザの指定した数だけ記事を抜き出す．これらはトップ記事として第一面に配置される．このとき新聞の紙面上に配置される記事には，記事の位置と大きさという 2 つのパラメータが存在する．

一般的な新聞のレイアウトでは，人の目線の移動に沿うよう順に，主要な記事が配置されている．そこで本インタフェースにおいても，重要であると思われる記事ほど，紙面の先頭に大きく配置する手法を採用する．つまり記事の 2 つのパラメータを，そのバースト度に応じて適宜変化させることで，柔軟なレイアウト構成を行う．

各記事には，記事のアスペクトと見出しが添えられて表示される．アスペクトは記事の観点を表すキーワード集合を意味し，このアスペクトを選択することで同じアスペクトを持つ Web ページ集めた詳細紙面に移ることができる．また見出しとは記事の概要を表す文章であり，Web ページから見出しとなる関係にあるコンテンツを抽出し，表示する．

4. 検索結果集合からの記事の抽出

検索エンジンに対してクエリを投げた時，返ってくる検索結果は指定したキーワードが含まれる Web ページの URL と，その Web ページ上の文章の断片の羅列である．これらの検索結果ページ集合の中には，同じかあるいは類似の内容を扱う文書が複数含まれている可能性がある．しかし，ユーザは実際にそれらのページのひとつひとつを開かないことにはその記事の内容を確認できず，何度も類似した内容のページを閲覧してしまう可能性がある．

このような検索結果閲覧の効率低下を避けるため，類似したトピックを持つ複数の文書はその類似度の基づいてクラスタリングし，それらをひとつの記事として扱うことを考える．

4.1 検索結果のクラスタリング

システムはまず Google Web Api [7] を使用し，ユーザの与えたクエリについて検索結果ページ集合を得る．

次に得られた各ページについて，ページ内の全文字数と，アンカー文字列を除く文字数の全文字数に対する比を計測し，これらの値がある閾値以上のページのみを 100 件抜き出すことで，クラスタリング対象文書とする．一般的に検索結果は Web のある時点におけるスナップショットであり，リンク切れページが含まれている可能性がある．こうしたリンク切れページや，リンク集ページなどのトピックに対する記述を多く含まないページを除外することが，この作業の目的である．

得られた各対象文書について，そのリンクアンカーを除いた文字列を形態素解析ツール茶筅 [8] で解析し名詞のみを抽出することで，$tf \cdot idf$ ベクトル空間における当該文書の特徴ベクトルを計算する．

最終的に，その特徴ベクトル間の類似度に基くクラスタリングを行うことで，得られた各クラスタを記事と定義し，抽出する．

4.2 記事の観点抽出

クラスタリングによって得られた各記事について，その記事の持つ観点を抽出することを考える．

Web 空間上に存在する文書の意味は，その文書自身が有する情報に加え，これらの文書の周辺にどのような文書が配置されているかによって推定することができる．是津ら [4] は，ある文書が他の文書からどのように参照されているのかを，この周辺情報から推定する手法を提案している．

一般的に，注目している対象ページへのリンクアンカーの周辺には，その対象ページに対する社会的評価が記述されていると考えることができる．リンクアンカーを辿り，対象ページに対して見出しの役割を果たしていると思われる見出しコンテンツを抽出し，クラスタリングする．是津らは，各クラスタに含まれる Web コンテンツから典型的に当該クラスタを表現するキーワード集合を，その対象ページに対するアスペクトと呼ん

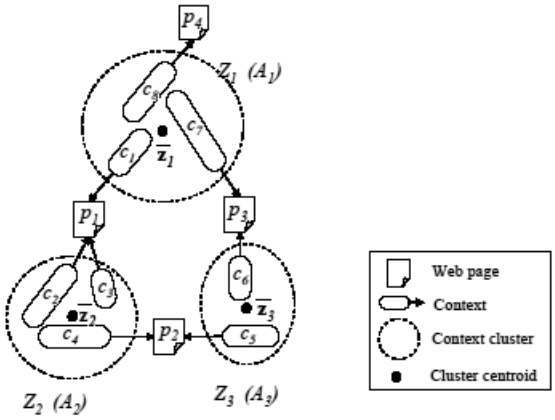

図 3　文書のアスペクト

でいる．

本研究では，各クラスタ内に存在するすべての文書に対しそのリンク元ページを求め，そこから得られる見出しコンテンツを当該文書のアスペクトとする．

まず記事のアスペクトを得るために，クラスタ内の対象文書へリンクしているリンク元ページ群を Google Web Api [7] の”link:”検索により特定する．これらリンク元ページの中から，外部サイトへのリンクを多数有するリンク集ページを求める．これは，一般にリンク集ページにおいて，リンクアンカー周辺やその上位階層には，対象ページに対する見出しが記述されていることが多いからである．このリンク集ページ中の見出しコンテンツを対象文書のコンテクストとし，抽出する．

最後に，得られた文書のコンテクストを，ベクトル空間モデルにおける特徴ベクトルで表現する．各特徴ベクトルの要素中から，大きな tf-idf 値を持つ単語を数個抜き出す．こうして得られたキーワード集合を，当該文書のアスペクトとして定義する．

5.　検索結果ページ集合のバースト度

新聞の第一面には普通，発行された時点において最も話題性の高い記事が配置される．検索結果ページ集合から抽出された記事群を対象に主要性に基づく順位付けを行う場合，単純に PageRank 値などをこの主要性の指標とするのは相応しくないと思われる．なぜなら PageRank 値は，注目している文書の被リンク数に基づいて算出されているため，時間の経過とは無関係に，常に人気の高い記事や一般的な内容の記事が上位にランキングされる傾向にあるためである．

そこで本研究では新聞の定期刊行性を利用し，記事を構成するクラスタの時間経過に伴う変化を調べることで，検索結果ページ集合のバースト度を定義する．第 4. 章によって抽出された記事集合の中から，新聞の第一面に掲載すべきトップ記事を選択する際，この検索結果ページ集合のバースト度を主要性の基準として使用する．

Kleinberg [1] は隠れマルコフモデルとポワソン分布を用いて，文書ストリームにおける文書の到達頻度を監視することによるトピック検出法を提案した．本研究では新聞が定期刊行物であるという点を利用し，文書ストリームを使わずに，ユーザがシステムに与えたクエリを定期的に検索して，その時間変化を調べることでバースト度を計算するという手法をとる．

今，t 番目に得られる (つまり t 番目に発行される新聞の) 検索結果ページ集合において，トピック i を記述する記事のクラスタを $C_{t,i}$ とする．クラスタ $C_{t,i}$ のバースト度を計算する際には，あらかじめ $C_{t,i}$ との類似度が最も大きな $t-1$ 番目のクラスタ $C_{t-1,i}$ を対応クラスタとして求めておく．このときクラスタ $C_{t,i}$ のバースト度は次の 3 つの要素から定義される．

5.1　クラスタ内の文書数に基づくバースト度

あるクラスタについて，その文書数が前回の検索時より増加していれば，そのクラスタが表すトピック i について，関連する新事実や著者の意見が書き加えられたと推測できる．つまりそれだけそのトピック i の主要性は高くなったと考えられる．図 4 はこのときのバースト度が上昇していくイメージである．

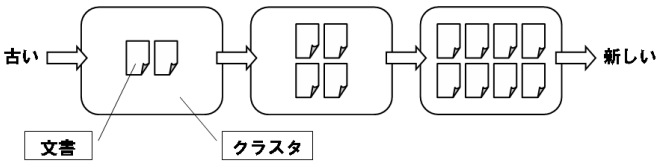

図 4　文書数に基づくバースト度のイメージ

クラスタ内の文書数の変化割合 $\Phi_{num}(C_{t,i})$ は次式で表される．

$$\Phi_{num}(C_{t,i}) = \frac{|C_{t,i}| - |C_{t-1,i}|}{|C_{t,i}| + |C_{t-1,i}|} \tag{1}$$

ただし，$|C|$ はクラスタ C 中に存在する文書数を表すものとする．この定義は，$\Phi_{num}(C)$ が必ず $-1 \leqq \Phi_{num}(C) \leqq 1$ を満たすことを保証する．

5.2　クラスタ内の文書の被リンク数に基づくバースト度

あるクラスタについて，その文書の被リンク数が前回の検索時より増加していれば，そのクラスタ内の文書を参照しているページが増加し，トピック i に注目する人が増加したと推測できる．つまりこのとき，ことトピック i の主要度は高くなったと考えられる．図 5 はこのときのバースト度が上昇していくイメージである．

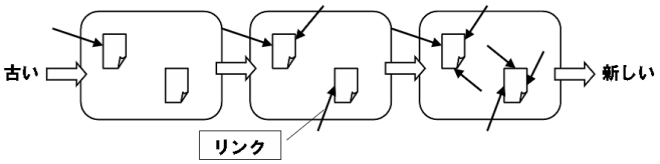

図 5　被リンク数に基づくバースト度のイメージ

クラスタ内の文書群の被リンク数の変化割合 $\Phi_{inlink}(C_{t,i})$ は次式で表される．

$$\Phi_{inlink}(C_{t,i}) = \frac{\sum_{p \in C_{t,i}} inlink(p) - \sum_{q \in C_{t-1,i}} inlink(q)}{\sum_{p \in C_{t,i}} inlink(p) + \sum_{q \in C_{t-1,i}} inlink(q)} \quad (2)$$

ただし，$inlink(p)$ はページ p の被リンク数を表すものとする．また，このとき必ず $-1 \leqq \Phi_{inlink}(C) \leqq 1$ となる．

5.3 クラスタの標準偏差ベクトルに基づくバースト度

クラスタ $C_{t,i}$ 内に存在する各文書の特徴ベクトルについて，ベクトル空間内における標準偏差ベクトルを計算し，その長さを求める．クラスタ $C_{t,i}$ の標準偏差ベクトルの長さがクラスタ $C_{t-1,i}$ の標準偏差ベクトル長さよりも短ければ，話題のばらつきが抑えられる方向に変化したと考えられる．このときこのトピック i は一つの方向性を持って収束していくものとし，トピックの主要度は高くなったとみなす．

また一方で，クラスタ $C_{t,i}$ の標準偏差ベクトルの長さがクラスタ $C_{t-1,i}$ の標準偏差ベクトルの長さより長ければ，このトピック i はその話題の構造に広がりをもって変化したものと思われる．しかし一方である程度話題が広がると，その話題は停滞するようになり，クラスタ内の文書は他のクラスタに吸収されてしまうはずである．そのためここでは，標準偏差ベクトルの長さが減少する場合においてのみ，高いバースト度を与えるものとする．図 6 はこのときのバースト度が上昇していくイメージである．

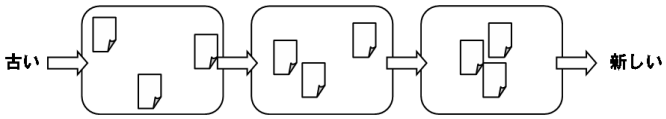

図 6　標準偏差ベクトルの長さに基づくバースト度のイメージ

クラスタ内の標準偏差ベクトルの長さの変化割合 $\Phi_{stdev}(C_{t,i})$ は次式で表される．

$$\Phi_{stdev}(C_{t,i}) = -\frac{stdev(C_{t,i}) - stdev(C_{t-1,i})}{stdev(C_{t,i}) + stdev(C_{t-1,i})} \quad (3)$$

ただし，$stdev(C)$ はクラスタ C の標準偏差ベクトルの長さを表すものとする．また，このとき必ず $-1 \leqq \Phi_{stdev}(C) \leqq 1$ となる．

5.4 クラスタの総合的なバースト度

以上 3 つの定義を用いて，クラスタ $C_{t,i}$ における検索結果ページ集合の総合的なバースト度 $\Phi(C_{t,i})$ を次式で定義する．

$$\Phi(C_{t,i}) = \alpha\Phi_{num}(C_{t,i}) + \beta\Phi_{inlink}(C_{t,i}) + \gamma\Phi_{stdev}(C_{t,i}) \quad (4)$$

ただし，α, β, γ はそれぞれ正の定数である．

またこのとき，クラスタ C のバースト度 $\Phi(C)$ は必ず次式を満たす．

$$-\alpha - \beta - \gamma \leqq \Phi(C) \leqq \alpha + \beta + \gamma$$

検索結果ページ集合から得られたすべての記事についてこのバースト度を計算し，トップ記事の抽出と記事の紙面上へのレイアウトを行う．

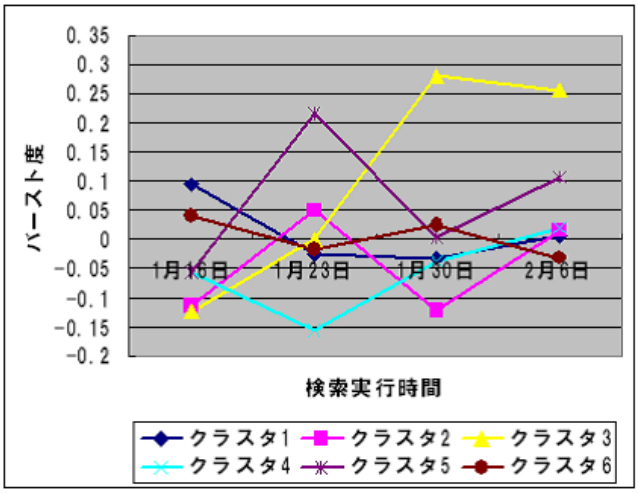

図 7　”スマトラ沖地震”による検索結果ページ集合のバースト度

5.5 実　　験

5.5.1 実験の目的と方法

バースト度の有用性を示すため，実際に検索エンジンにクエリを与え，その検索結果ページ集合の時間的変化を評価する実験を行った．検索エンジンとして Google Web Api [7] を使用し，検索エンジンに与えるクエリは「スマトラ沖地震」で，2005 年 1 月 9 日から 2005 年 2 月 6 日まで 1 週間おきに合計 5 回検索を実行した．これらから得られる各記事についてそのバースト度を計算し，その時間的な推移を調べた．

図 7 は，各検索の時点で得られる検索結果を 6 つのクラスタへクラスタリングし，その各クラスタのバースト度を時系列順に折れ線グラフで表現したものである．

また，表 1 は 2005 年 1 月 9 日の，表 2 は 2005 年 2 月 6 日の検索結果について，それぞれクラスタの文書数・被リンク数・標準偏差ベクトルの長さ，およびクラスタ内に生起する単語のうち，大きな重みを持つ単語から 10 個をキーワードとして抜き出してきたものである．この実験において，上記に示した 3 つのバースト度の重みの値として，それぞれ $\alpha = 0.3, \beta = 0.1, \gamma = 0.3$ を用いた．

5.5.2 実 験 結 果

図 7 を見ると，クラスタ 1，クラスタ 2，クラスタ 4，クラスタ 6 のバースト度が，各検索結果を通して比較的低い値となっていることがわかる．一方でクラスタ 3 のバースト度は当初から徐々に増加し，1 月 23 日頃にピークを迎えている．またクラスタ 5 のバースト度は 1 月 16 日頃に急激に上昇し，その後また減少している．

ここで，クラスタ 1 は，そのキーワードとして ”募金 ”や ”義援金 ”が現れることから，被災地への援助を呼びかける記事を構成することが推測できる．実際，クラスタ 1 内に存在する文書には，ユニセフの緊急支援情報 [9] などのホームページが見られた．また同様にしてクラスタ 2 にはチャリティなどの情報が，クラスタ 3 には現地での衛生状況の悪化を伝える情報が見られ，クラスタ 4 にはスマトラ沖地震に対する被害状況などの情報が，クラスタ 5 には地震の規模を伝える情報などが，ク

クラスタ名	文書数	被リンク数	標準偏差ベクトルの長さ	キーワード
クラスタ 1	45	2365	0.0075568	寄付 募金 円 義援金 手数料 リリース 免除 送金 口座 アジア
クラスタ 2	13	51	0.0051022	円 コンテンツ チャリティ 件 会員 バックアップ 面 旅行 販売
クラスタ 3	2	0	0.0015232	検疫 仮設 建設 合同庁舎 住居 世帯 村 住宅 感染 港湾
クラスタ 4	20	16	0.0054511	人 香川 死者 日本人 死亡 案内 新着 数 不明 火
クラスタ 5	9	13	0.0035225	画像 衛星 受信 地球 結果 の 投稿 検出 球形 自転
クラスタ 6	11	148	0.0040028	久家 玉名 隊 債務 G7 NPO 撤退 人 調査 米

表 1　2005 年 1 月 9 日 ”スマトラ沖地震”の検索結果上位 100 件のクラスタリング結果

クラスタ名	文書数	被リンク数	標準偏差ベクトルの長さ	キーワード
クラスタ 1	46	8284	0.0070146	募金 円 義援金 口座 手数料 送金 免除 ジャパン 寄付 受付
クラスタ 2	14	4490	0.0057355	ダイヤル チャリティ 運航 チャンネル 個 月 料 航空 年
クラスタ 3	6	134	0.0034235	分 ライブ ドア 青少年 時 福祉 検疫 たち 精神 子ども
クラスタ 4	12	3772	0.0050166	不明 人 行方 カ月 死者 共通 未確認 日程 同日 平田
クラスタ 5	13	1095	0.0065131	地球 画像 自転 NASA 変化 検出 結果 衛星 球形 コメント
クラスタ 6	9	674	0.0039186	人 朝日新聞 防衛庁 米 撤退 自衛隊 米国 労働 派遣 部隊

表 2　2005 年 2 月 6 日 ”スマトラ沖地震”の検索結果上位 100 件のクラスタリング結果

ラスタ 6 には調査隊派遣に関する情報などが多く見られた．

5.5.3　評価と考察

- クラスタ 1 内に多く存在する義援金や援助呼びかけのページなどは，スマトラ沖地震発生直後から検索結果の上位にランクされ，その後も安定して上位に位置する傾向にある．このためにクラスタ 1 のバースト度は，増加しなかったのではないかと推測できる．
- クラスタ 4 内に多く存在するスマトラ沖地震の被害状況そのものを伝えるページなどは，1 月 9 日の時点で既に多く存在していた．その後このクラスタ内の文書数は減少していったため，バースト度は増加しなかったのではないかと思われる．
- クラスタ 3 は現地における衛生状況の悪化などに関する記事を構成するが，このクラスタの文書数は検索の初期段階において非常に少ない．しかしその後次第に文書数は増加し，そのために標準偏差ベクトルの長さは大きくなったが，同時に被リンク数も増加したために結果としてバースト度は増加したことがわかる．
- クラスタ 5 には地震の規模を伝えるページが多く存在する．そこへ，スマトラ沖地震の影響で地球の自転速度がわずかに速まったことを伝えるニュース記事が，1 月 11 日付けで発生した．クラスタ 5 のバースト度が 1 月 16 日の時点で上昇しているのは，このニュース記事に対するリンク元ページがその後急激に増加したためではないかと思われる．
- 6 つのクラスタすべてについて，被リンク数の急激な増加が見られた．Web ページの被リンク数はハイパーリンクの性質から普通，時間とともに増加するものであるが，ここまで急激な増加が起きたのは，検索結果ページ集合にニュースサイトのページが多く含まれていたためであると考えられる．というのも，ニュース記事中には関連記事として他の記事へのリンクが多く含まれている場合があるからである．そこで今回の実験では，この被リンク数の影響を抑えるため，被リンク数に基づくバースト度の重み β を小さく設定した．
- 実験では，クラスタ対象ページを検索結果より 100 件取得したが，そのほとんどがクラスタ 1 に分類されてしまったため他のクラスタの文書数が少なく，少しの変化により，そのバースト度が影響されやすくなってしまった．

以上の考察より，地震発生から時間が経過するとともに，地震の被害や支援・援助に関する情報から，衛生状況の悪化や地震の規模といった情報に人々の関心が移っていったことが読み取れる．これらの点で，バースト度はうまくトピックを検出できたと思われる．

また，検索結果ページ集合のバースト度は，ページ集合のクラスタリング結果に大きく左右されることがわかった．実際，”イラク国民議会選挙 ”というキーワードで検索を実行すると，その検索結果ほとんどが 2005 年 1 月 30 日に選挙が開催されるという事実を述べる Web ページであるため，クラスタがほとんど 1 つにまとまってしまった．このような場合，少ない文書数からなるクラスタのバースト度は，他のクラスタのバースト度に比べ，変動しやすくなってしまうものと考えられる．

6.　新聞インタフェース

3. 章，4. 章，および 5. 章で説明した手法を実装し，検索結果からバースト度に基づき検出されたトピックの動向・変遷の把握を支援するためのシステムとして，新聞インタフェースのプロトタイプを構築した．

6.1　プロトタイプ実装

図 8 は実際に新聞インタフェースにより，取得された検索結果を閲覧している様子である．ユーザはあらかじめシステムに対して，関心を持つトピックについてのキーワードと表示される記事数（紙面の分割数）を与える．その後，アプリケーションから検索を実行させ，得られた検索結果と前回の検索結果との比較を行い，そのバースト度に基づいて新しい新聞をインタフェース上に表示する．

図の例では 5 つの記事が新聞インタフェース上に配置され，それぞれそのバースト度ごとに大きさや配置位置といったレイアウトが異なっている．またひとつの記事はアスペクト・見出

図 8　新聞インタフェースの外観

し・本文の 3 要素から構成される．

アスペクトとしては，記事の観点を表現するキーワード集合から，特徴量の大きいものを 2 つ抜き出して表示してある．ユーザはインタフェース上のアスペクト欄をクリックすることで，実際の新聞において ”経済面 ”や ”スポーツ面 ”といった詳細紙面へと新聞をめくる様に，同一アスペクトからなる Web ページのみを集めた詳細紙面に移動することができる．これらの詳細紙面においても，第一面と同様に各記事はバースト度に基づいてランク付けされ，それに基づいてレイアウトされている．

また新聞インタフェース上に表示される記事の本文には，当該トピックを表現するクラスタ内の Web ページのうち，その特徴ベクトルとクラスタの特徴ベクトルの類似度が最も高いページを表示している．

記事の見出しには，表示しているページ中から，title や h1, h2 などの見出しの役割を果たすタグに囲まれた見出しコンテンツを選択し，表示する．

6.2　考　　察

今回のプロトタイプにおいては，記事の紙面としてクラスタ内の特定のページをそのまま表示しているため，決して新聞らしい表示がなされていないという問題点が挙げられる．このために，クラスタ内の文書間における差異が適切に表現されないので，新聞の網羅性としての性質は小さくなってしまったのではないかと考えられる．

さらに，実際の記事の場合を考慮すると，記事はひとつだけではなく複数の観点に属することがある．例えば,「スマトラ沖地震における被災地援助のため，日本が 5 億ドルの資金援助」という記事は，経済面や国際面などの複数の観点を持つ．こうしたケースを今回実装したインタフェースは無視してしまっているため，今後これらをうまく表現する余地が残されている．

また今回提案した手法では，システムが保持する過去の検索結果系列に対して，ひとつ前の検索結果との比較のみを行っている．今後，検索結果系列全体との比較をすることでさらに良い精度で，トップ記事を抽出できるのではないかと思われる．

7.　ま　と　め

本論文は，検索結果ページ集合のバースト度を定義し，クラスタリングと観点抽出に基づく閲覧インタフェース提案した．またそのプロトタイプを実装し，トピックの変遷が把握できるか実験を行った．本論文で提案する新聞インタフェースによって，従来の検索エンジンだけでは困難であった動向や変遷の把握と，検索結果の一覧表示が可能になる．

今後の課題として，上記に挙げた新聞インタフェースの改良，及びバースト度の精度向上といった問題点の修正に取り組みたいと思う．また異なるメディア間でのバースト度測定と，その比較を行うことも考えていきたい．

謝　　辞

本研究の一部は，文部科学省科学技術振興費プロジェクト「異メディア・アーカイブの横断的検索・統合ソフトウェア開発」（代表：京都大学田中克己）および，平成 16 年度科研費特定領域研究 (2) 「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」（課題番号：16016247，代表：田中克己）および，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」によるものです．ここに記して謝意を表すものとします．

文　　献

[1] J.Kleinberg, ”Bursty and Hierarchical Structure in Streams”, In *Proc. ACM SIGMKDD*, 2002
[2] National Institute of Standards and Technology: *http://www.nist.gov/speech/tests/tdt/*
[3] G.Golovchinsky, M.H.Chignell and L.Rostek, ”VOIR: visualization of information retrieval”, In *ACM SIGWEB Newsletter*, 1996
[4] Zettsu, K., Kidawara, Y. and Tanaka, K, ”Aspect Discovery: Web Contents Characterization by Their Referential Contexts”, In *APWeb '04*, 2004
[5] J.Cho and S.Roy, ”Impact of Web Search Engines on Page Popularity”, In *World-Wide Web Conference*, 2004
[6] Google: *http://www.google.co.jp/*
[7] Google Web APIs: *http://www.google.com/apis/*
[8] 奈良先端科学技術大学松本研究室　形態素解析ツール茶筌: *http://chasen.aist-nara.ac.jp/hiki/ChaSen/*
[9] unicef スマトラ沖地震・津波情報ページ: *http://www.unicef.or.jp/kinkyu/sumatra/2004.htm*