特集　Webダイナミクス

6 Web情報ストリーム

高間　康史　東京都立科学技術大学工学部電子システム工学科／PRESTO（科学技術振興事業団）
ytakama@cc.tmit.ac.jp

オンラインニュースや掲示板，メールマガジンなど，Web上で発信・やりとりされる情報の中には互いに時系列的関連を持ち，情報ストリームを形成するものが少なくない．これらの情報ストリームから，新規話題の発生を自動検出したり，興味ある話題の関連記事を自動追跡・収集したりすることができれば，ともすれば情報過負荷に陥りがちなWeb情報活用において非常に強力な武器になることが期待できるが，関連技術・研究はまだ多くない現状にある．

本稿では，Web情報ストリームを扱う技術として，話題検出・追跡技術に関するTDT，および情報ストリーム可視化技術について最新の事例を中心に紹介する．

Web情報ストリームとは

情報ストリームと言われてすぐに思い浮かぶのは，映像ストリームであろう．しかし，広義に「データ・情報が時系列的関連を持って並べられたもの」と定義すれば，我々の身近なところでいろいろ見つけられることに気づく．我々人間が時間の概念を持ち，時間軸に沿って生活していることを考えれば，情報・データの持つストリーム性は本質的なものであるといっても過言ではないであろう．

特に，計算機の進歩や常時接続環境の普及などにより，その進歩・拡大が著しいWebにおいては，情報発信および蓄積･収集の容易さから，大量の情報が日々出現し，その中にはオンラインニュースや掲示板，メールマガジンなど，情報ストリームを形成するものも少なくない．たとえばWeb上では，新聞社，雑誌社を始めとする数多くのサイトにおいてオンラインニュースが公開されており，日々あるいはもっと頻繁に記事をアップデートする．これらのニュース記事を情報ストリームとして扱い，新規話題の発生を自動検出したり，興味ある話題の関連記事を自動追跡・収集したりすることができれば，ともすれば情報過負荷に陥りがちなWeb情報活用において非常に強力な武器になることが期待できる．

本稿では，Webにおいて発見される，上述のような時間的関連を持った情報群をWeb情報ストリームと呼ぶ．Web情報ストリームの重要性については，近年指摘されるようになってきたものの，関連研究はまだそれほど多くはない．ここでは話題検出・追跡技術に関するTDT，および情報ストリーム可視化技術について最新の事例を中心に紹介する．

TDT: 情報ストリームからの話題検出・追跡技術

Topic Detection and Tracking （TDT）とは

Topic Detection and Tracking（TDT）（http://www.nist.gov/speech/tests/tdt/）は，オンラインニュースやニュース放送といったデータストリームから話題構造を自動的に得るための技術確立を目的として，DARPAにより進められている技術研究開発プログラムである[1)]．1997年にパイロットスタディが開始された後，2回目のTDT1998より，その評価がオープンとなり，TDT1999, TDT2000, TDT2001と開催され，現在は5回目のオープン評価（TDT2002）の最中である．

TDTの特徴は，TREC（Text REtrieval Conference）などと同じく，共通のコーパス，タスク，評価基準を用意することにより，システム・手法間の比較検討を可能にし，関連技術の発展を促進する点である．必ずしもWeb情報ストリームを対象としているわけではないが，情報検索システムでもおなじみの，比較的汎用性の高い手法

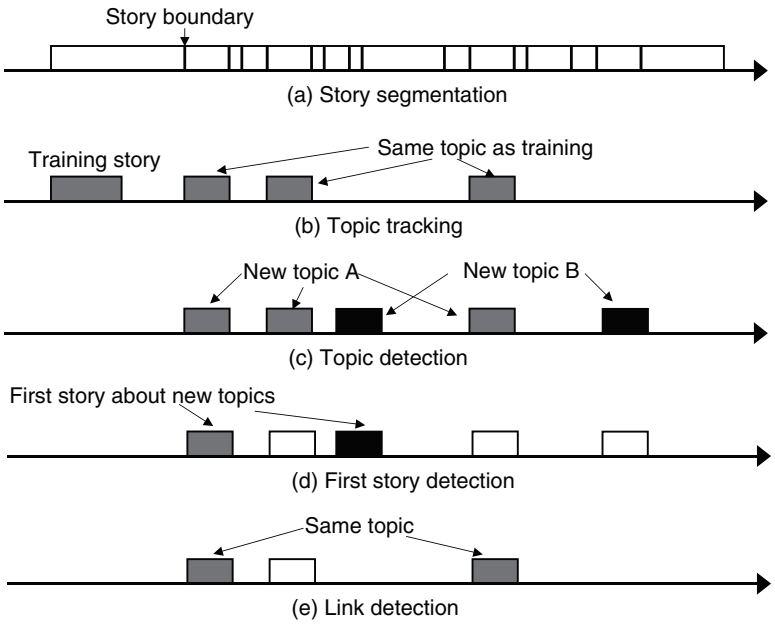


図-1　TDTの5つのタスク

が用いられることが多いため，Webとの親和性は高いと考えられる．特に，オンラインニュースはWeb情報ストリームの代表的なものであり，情報抽出ラッパーなどと組み合わせることにより容易に適用可能であろう．以下ではTDTの概略を述べた後，代表的タスクとして話題検出・追跡タスクについて詳しく紹介する．

TDT評価システム

■コーパス

TDTでは，情報ソースやメディア，領域，言語，アプリケーションに依存しないアルゴリズムの確立を目的としており，これを考慮してコーパスも複数のソース，メディア，言語から構成されている．具体的には，Webやオンラインニュースから収集されたテキストデータと，ラジオやテレビから収集した音声データを含んでいる．それぞれ，英語と中国語（マンダリン）のデータがあり，中国語については英訳したものも一部用意されている．音声データについては，オーディオデータ（audio sampled data signal），人手による転写（manual transcription），ASR（automatic speech recognition）による自動転写の3種類が用意されている．

評価や学習のために，各記事について話題ごとにYES，NO，BRIEFのタグがつけられている．その話題についての記事である場合はYES，関係ない場合はNOであり，その記事に関する記述が記事中の10%に満たない場合はBRIEFとタグ付けされる．

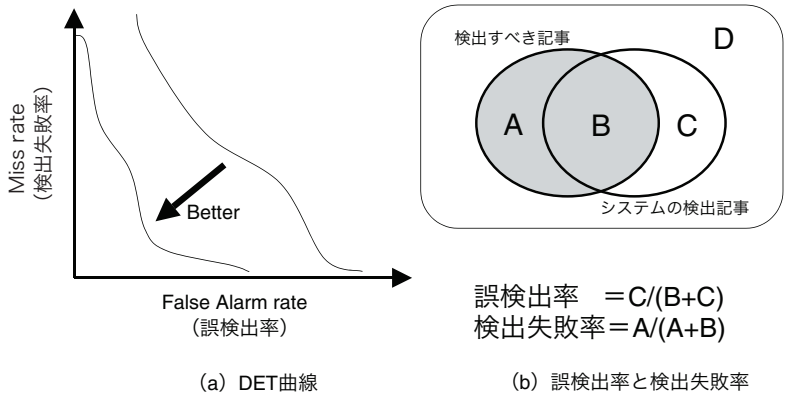
現在，3種類のコーパスTDT-Pilot, TDT2, TDT3がLDC（Linguistic Data Consortium）において公開されており，TDT2002における評価のためにTDT4が作成中である．TDT-Pilotコーパスは1994年7月～1995年6月の間にReuterやCNNから収集された約26,000記事のテキストデータから構成され，25話題についてタグ付けされている．TDT2コーパスは1998年1～6月に収集された約57,000記事（100話題）からなり，音声データも含まれる．TDT3コーパスは，1998年10～12月に収集された約45,000記事（240話題）からなるが，TDT2001において1998年7～9月のデータも含むように拡張され，約80,000記事を含むようになった．

■タスク

TDTではオンラインニュースなどから収集した，時間順序に並んだ言語データセット（コーパス）を対象として，以下の5つのタスクを設定している（**図-1**）．

- 記事分割（story segmentation）：ある出来事について述べた，話題的に同質な断片（記事）の発見（図-1 (a)）
- 話題追跡（topic tracking）：与えられた話題に関する記事の発見（図-1 (b)）
- 話題検出（topic detection）：新規話題の発見，関連記事のクラスタリング（図-1 (c)）
- 第一記事検出（first story detection）[☆1]：新規話題の第

☆1　第一記事検出はTDT2002より新規イベント検出（new event detection）と名称が変更されている．



(a) DET曲線　　(b) 誤検出率と検出失敗率

図-2　TDTの評価基準とDET曲線

タスク	P_{target}	C_{Miss}	C_{FA}
記事分割	0.3	1.0	0.3
話題追跡	0.02	1.0	0.1
話題検出	0.02	1.0	0.1
第一記事検出	0.02	1.0	0.1
リンク検出	0.02	1.0	0.1

表-1　コストパラメータのタスクごとの設定

一記事発見（図-1 (d)）

- リンク検出（link detection）：記事が同一話題に属するかどうかの判定（図-1 (e)）

第一記事検出と話題検出は非常に類似したタスクであり，ある新規話題に関して，第一記事検出は時間軸上で最初に出現する記事のみを検出するのに対し，話題検出では以降の関連記事すべてを検出し，その話題に対応する記事クラスタを生成するまでがタスクとなる．

TDTにおける話題とは，「ある特定の時間・場所において発生した出来事」および，それと直接に関連する出来事，活動として現在は定義されている☆2．Fiscusらの例によると[2]，ある航空機墜落事故が起きた場合，生存者の捜索や犠牲者の葬儀などは直接関連する出来事と見なされ，同一話題に含まれる．一方，米連邦航空局（FAA）が墜落調査の結果示した修理命令などは直接関連がないとして話題には含まれない．もちろん，他の航空機事故はまったく別の話題として扱われる．

■評価基準

各タスクの評価基準として，**図-2**（a）に示すDET（Detection Error Tradeoff）曲線が採用されている．これは，横軸にとった誤検出率（false alarm rate）と縦軸にとった検出失敗率（miss rate）との間のトレードオフを示したものである．話題追跡タスクを例にとると，誤検出率は指定話題に関する記事とシステムが判断したもののうち，関連しない記事の割合，検出失敗率は，指定話題に関する記事のうち，システムが検出できなかった記事の割合である（**図-2**（b））．情報検索分野の代表的尺度と比較すると，高適合率（precision）は低誤検出率，高再現率（recall）は低検出失敗率となる．したがって，情報検索システムにおける適合率—再現率曲線はグラフの右上を通るほどよいが，DET曲線は原点に近づくほど高性能となる．

また，統合的指標として，正規化コスト（normalized cost）が式（2）で定義される．

$$C_{Det}=C_{Miss}\cdot P_{Miss}\cdot P_{target}+C_{FA}\cdot P_{FA}\cdot\left(1-P_{target}\right), \quad (1)$$

$$(C_{Det})_{Norm}=C_{Det}/\min\left(C_{Miss}\cdot P_{target}, C_{FA}\cdot\left(1-P_{target}\right)\right). \quad (2)$$

ここで，C_{Miss}，C_{FA}は検出失敗，誤検出の際にかかるコスト，P_{Miss}，P_{FA}は検出失敗率，誤検出率，P_{target}は対象データにおいて正答（検出すべき記事など）が出現する確率を表す．パラメータの値はタスクごとに決定され，TDT2001では**表-1**の値が用いられている．

話題追跡タスク

話題追跡タスクでは，ある特定の話題に関する少数の記事（1～4記事）が与えられた場合に，その話題に関する記事をテストデータストリームから検出する．このとき，与えられた訓練記事からクエリーあるいはプロファイルを生成し，類似度があるしきい値以上の記事を関連記事として検出する問題と定式化すれば，情報検索システムと同様のアプローチをとることができる．TFIDFやベクトル空間モデルも利用されているが，大量のコーパスが利用可能であることから統計的アプローチをとるシステムも多い．TDT2001で好成績を残したLIMSIのシステム[3]（以降LIMSI2001と呼ぶ）ではunigram modelを採用しており，文書dと話題Tの類似度$S(d, T)$を正規化対数尤度比（normalized log likelihood ratio）で定義し（式（3）），この値がしきい値（th_D）を超える記事について，与えられた話題に関連していると判断する．

☆2 1997年のパイロットスタディでは，特定の時間・場所において発生した出来事に限定されていた．

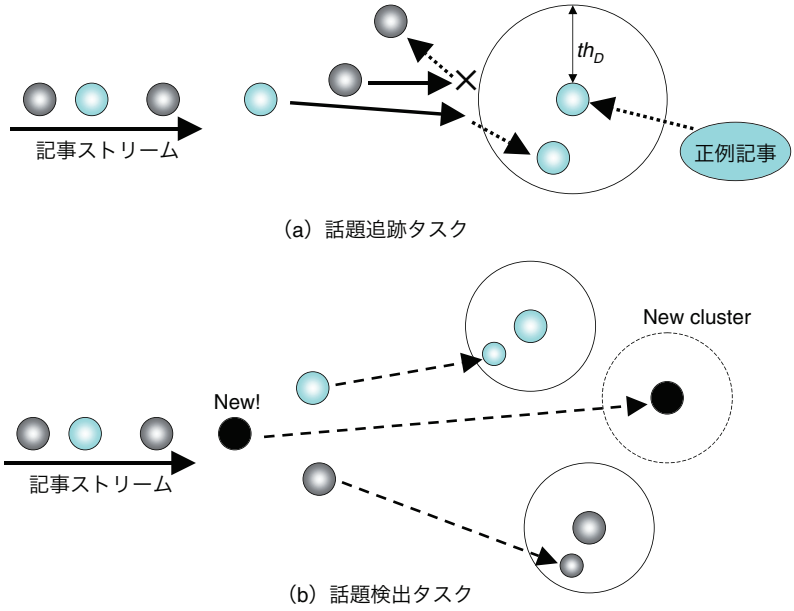


図-3 話題追跡タスクと話題検出タスクの比較

$$S(d, T)=\frac{1}{L_d}\sum_{w\in d} tf(w, d)\log\frac{\lambda P(w|T)+(1-\lambda)P(w)}{P(w)} \quad (3)$$

ここで，L_d は d の長さ，$tf(w, d)$ は単語 w の d における出現頻度，$P(w|T)$ は話題モデル T が与えられた場合の w の生起確率に関する最尤推定量，$P(w)$ は w の生起確率に関する背景モデル（general English model. 訓練用コーパス（TDT2）から計算）である．

$P(w|T)$ は少数記事から求められるため，標本値のスパースさが問題となるが，式（3）では $P(w)$ を用いて補間することにより，この問題に対処している．補間係数 λ の値は訓練用コーパスにおいて追跡コストを最小とする値（$\lambda=0.25$）を用いている．少数の正例記事から求めた単語の生起確率を大量の訓練コーパスから求めた統計量を用いて補正する方法はスムージング（smoothing）とも呼ばれ，他のシステムでも導入されている．

訓練例の不足を補うアプローチはほかにもあり，LIMSI2001 ではストリームデータを処理しつつ訓練データとして利用可能な関連記事を取り込んでいく，教師なし適応（unsupervised adaptation）も利用されている．具体的には，関連記事の中でもより信頼性の高い記事のみを選択するために，検出に用いるしきい値（th_D）とは別のしきい値 $th_A(\geq th_D)$ を用意し，$S(d, T)$ がこれを超える記事のみを用いて単語出現頻度情報を更新する．訓練データの補充は TDT タスクにおいて重要であり，教師なし適応のアイディアは他のシステムでも利用されている．

他に統計量を利用したものとしては，Bayes 則を利用した確率モデルを用いるシステムや，単語の生起確率を β 二項分布（beta-binomial distribution）に基づいてモデル化するシステムなどがある．

話題検出タスク

話題追跡タスクと話題検出タスクでは類似点も多いが，後者は検出すべき話題に関する訓練データが与えられない点で異なる．すなわち，**図-3**（a）に示すとおり，話題追跡タスクでは与えられた訓練データを基に検出すべき話題に対応するクラスタを事前に生成し，時系列に入力される記事の中でクラスタ内に含まれる（類似度がしきい値以上）ものを関連記事として検出する．これに対し話題検出タスクでは，事前にそのようなクラスタを生成することはできないので，入力データを処理しつつ類似記事のクラスタを漸次生成していく必要がある．新規話題の検出は，新たなクラスタの生成としてとらえられる．話題検出タスクを行う TNO2001 システム [4] の概要は以下の通りである．

1. 遅延ウィンドウ（deferral window）内の各記事 d について，既存クラスタの内で最も類似するもの $C_M(d)$ を見つけ，その類似度 $S_M(d)$ に基づき以下の処理を行う．
   A）$S_M(d) > th$（しきい値）→ d を $C_M(d)$ に割り当てる．
   B）それ以外 → d を基に新規クラスタを生成する．
2. 遅延ウィンドウの終わりまで到達したら，ウィンドウ内の各記事について，再び既存クラスタとの類似度を再計算し，最も類似するクラスタ $C'_M(d)$ を見つけ（類似度 $S'_M(d)$ とする），以下の処理を行う．
   A）$S'_M(d) > \max(S_M(d), th)$ → $C'_M(d)$ に d の割り当てを

変更する．

B）max $(S_M(d), S'_M(d)) < th$ → 新規クラスタを生成する．

C）上記以外 → クラスタ割り当てを変更しない．

話題検出タスクにおいては以降に続く記事のうち，一定数（10記事）を処理するまで話題検出の判断を遅らせることが許されているため，これに合わせて遅延ウィンドウを設定し，ウィンドウ内の記事を処理した後でクラスタの再配置を行っている（ステップ2）．ステップ1のシングルパス・クラスタリングで生成されるクラスタは，データの与えられる順序に依存するため，遅延ウィンドウを利用した再配置を行うことで，クラスタリング精度の向上を実現している．クラスタの再配置は，BBNやDragon Systemなど，他のシステムでも採用されている．

話題追跡タスクと話題検出タスクのもう1つの大きな違いは，検出タスクでは考慮すべき話題クラスタが複数あるため，計算コストが余計に必要な点である．通常のシングルパス・クラスタリングでは，データが追加されるたびにクラスタモデル（クラスタ中心など）を再計算する必要があり，計算コストが負担となる．TNO2001では入力記事とクラスタの類似度を，記事とクラスタ中の各記事との類似度に基づいて決定する方法を採用し，クラスタモデル再計算の負担を低減している．各記事間の類似度の計算については，式(3)と類似の正規化対数尤度比に基づいている．スムージング処理については，係数λに相当する部分が一定値ではなく，記事長に依存するDirichletスムージングを採用しており，短い記事が多い場合にはこちらの方が性能向上が期待できるとしている．

情報ストリームの可視化技術

映像ストリームなどは，それ自体は人間にとって可読なものであり，発見された構造（シーン，ショットなど）は見たい映像へのアクセスを容易にするメタ情報として利用される．これに対し，本稿で対象とするWeb情報ストリームは，そのままでは我々にとって分かりやすいものとはいえず，何らかの情報可視化手法が必要であるといえる．TDTでは話題構造の抽出が目的であり，発見・抽出した話題情報をいかに提示するかについては現在のところ考慮されていない．また，情報検索結果やリンク構造の可視化を行うシステムについてはその重要性が早くから指摘され，さまざまに提案・開発されているが，話題の発生・持続・消滅などの時間的変化を可視化するシステムについてはまだ少ない．ここでは川の流れのメタファを用いた話題ストリーム可視化システムThemeRiver[5]と，キーワードマップをベースとした可視化システム[6]を紹介する．

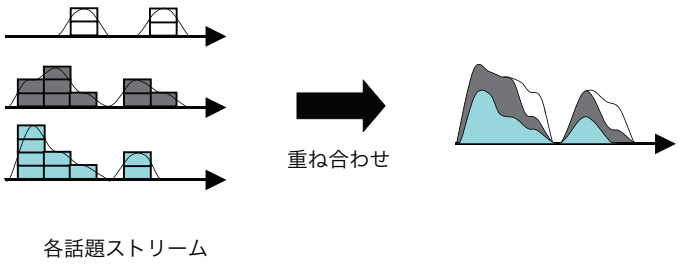


図-4 ThemeRiverによる可視化の概要

ThemeRiver

ThemeRiverは，大規模文書集合から話題の時間的変化を検出･可視化することを目的としたシステムであり，話題の変化，持続性，共起性などの時間軸に沿った変化を，川の流れのメタファで可視化する．具体的には，以下のメタファによって可視化する．

- 文書集合の時間的発展 → 川の流れ
- ある時点における話題の強さ（関連記事の多さ） → 川の幅
- 個々の話題を色分けし，重ね合わせて全体を表現．

図-4は，ThemeRiverの可視化の概要を示している☆3．情報ストリームは月単位で集計され，t_i月におけるj番目の話題に関するキーワードを含む記事数をy_{ji}として，$(t_i, y_{ji})(i=1, ..., n, j=1, ..., m)$の形式で表される．可視化する際には，以下の条件を満たす曲線$\hat{f}_j$で，各データ点間を補間する．

- $\hat{f}_j(t_i)=y_{ji}$.
- $\hat{f}'_j(t_i)=0$.
- $s \in (t_i, t_{i+1}), \min\left(y_{ji}, y_{j(i+1)}\right) \leq \hat{f}_j(s) \leq \max\left(y_i, y_{j(i+1)}\right)$

1番目の条件は，補間曲線が実際のデータを忠実に反映することを要求している．また，2, 3番目の条件を満たす曲線は，最終的に複数の話題の流れを重ね合わせたときにオーバラップすることがない．

☆3 図はあくまで概要を示したものであり，以下に示すアルゴリズムで厳密に求めた曲線ではない．

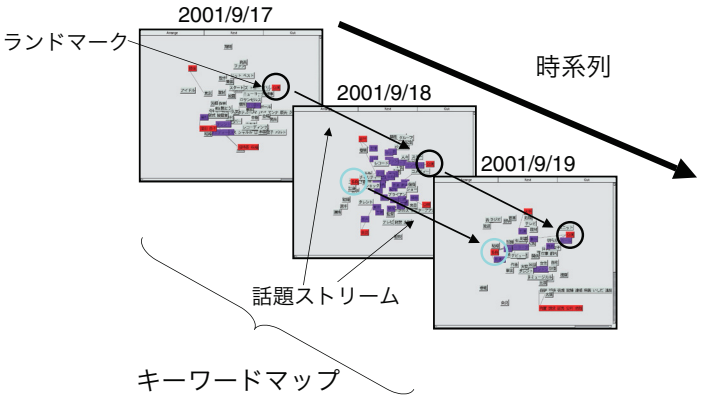


図-5　キーワードマップを用いた話題ストリーム可視化の概要

文献5）には，TREC3の配布ディスクに含まれているAP通信オンラインニュース記事（1990年）を可視化した例が示されており，関連記事数のヒストグラム表示と比較した結果，マクロな変動をとらえるのには有効であることを示している．

キーワードマップに基づく話題ストリーム可視化システム

高間らは，文書集合中の話題分布を抽出キーワードの2次元配置（キーワードマップ）により可視化する際に，読解の手がかりとなるランドマークキーワードを免疫ネットワークモデルに基づいて抽出・提示するシステムを提案しているが，オンラインニュース記事を発行日ごとにキーワードマップを用いて可視化し，これを時系列に並べることによって話題ストリームを可視化することを試みている（**図-5**）．各文書集合を可視化する際に，完全に独立に可視化したのでは時系列的関連性が失われやすいことから，以前に処理した文書集合からランドマークを考慮して可視化することにより，話題ストリームを優先的に抽出している．ニューヨーク同時多発テロ事件直後のオンラインニュースを可視化した例では，チャリティコンサートや，テロの影響によるコンサートのキャンセルなどに関する記事が多く見られ，話題ストリームが形成されていることなどを示している．TDTにおける話題のような特定の出来事に関する狭義なものではなく，より広義なものとして話題を扱うことにより，おおまかな話題の流れをとらえようとしたシステムといえる．

今後の展望

オンラインニュースなど，Web上では日々新しい話題に関する文書が登場し，情報ストリームを形成している．これらから新規話題を発見したり，興味ある分野における流行や変化などを検出・可視化することができれば，我々にとってWebはますます欠かせない情報源となるであろう．今回紹介したTDTや情報ストリーム可視化に関する研究はまだまだ歴史も浅く，限定された範囲での扱いに目処がついたに過ぎない．今後は，話題の粒度に関する検討や，話題間の関連や話題自身の変容をいかに扱うか，なども検討の対象になろう．

また，サーチエンジンを用いた情報検索は，キーワードを一度入力しただけで所望の結果が得られることは非常にまれであり，得られた検索結果を基に次回クエリーとして入力するキーワードを吟味し，検索を行うといったように，何回か検索を繰り返すのが通常である．これら一連の情報検索プロセスによって得られる検索結果集合も，ユーザ固有のものではあるが時間的関連を持った情報ストリームと見なすこともできる．特にユーザにとって未知の対象を検索する場合においては，この情報ストリームから対象分野の主要な話題や，検索プロセスの初期においてユーザが見落とした話題を後で指摘することなどができれば，非常に強力な支援となり得る．

Web上に蓄積される情報はますます増加する一方であり，情報ストリームが潜在的に含む情報もこれに伴い豊かになっていくことは間違いない．より長期のストリームが蓄積されるようになれば，工学の分野や最先端ビジネスの場だけでなく，社会学，歴史学的にも興味深いものとなっていくのではないだろうか．これからの発展が非常に期待される分野であるといえよう．

参考文献

1）Wayne, C. L.: Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation, Language Resources and Evaluation Conference (LREC) 2000, pp.1487-1494 (2000).
2）Fiscus, J. G. and Doddington, G. R.: Chapter 2: Topic Detection and Tracking Evaluation Overview, in J. Allan ed., Topic Detection and Tracking-Event-based Information Organization-, Kluwer Academic Publishers, pp.17-31 (2002).
3）Lo, Y. Y. and Gauvain, J. L.: The LIMSI Topic Tracking System for TDT2001, Proc. 2001 Topic Detection and Tracking (TDT) Workshop (2001).
4）Spitters, M. and Kraaij, W.: TNO at TDT2001: Language Model-Based Topic Detection, Proc. 2001 Topic Detection and Tracking (TDT) Workshop (2001).
5）Harve, S., Hetzler, E., Whitney, P. and Nowell, L.: ThemeRiver: Visualizing Thematic Changes in Large Document Collections, IEEE Trans. on Visualization and Computer Graphics, Vol.8, No.1, pp.9-20 (2002).
6）Takama, Y. and Hirota, K.: Web Information Visualization Method Employing Immune Network Model for Finding Topic Stream from Document-set Sequence, New Generation Computing, Vol.21, No.1, pp.49-59 (2002).

（平成15年4月1日受付）