社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－DBS－134 (II) (96)
2004／7／15

Webからの時制クラスタの解釈

森 正輝† 三浦孝夫† 塩谷 勇††

本稿では、Webページ集合からの事象抽出及び自動解釈を行うための新しいWebマイニングの手法を提案する。Webページを調査し有効時間を抽出しK-meansクラスタリングにより事象の抽出を行う。TDTのアプローチをWeb環境に適応し、提案する手法が時制Webページに対して有効であることをいくつかの実験により示す。

Labeling Temporal Cluster of Web Pages

MASAKI MORI,† TAKAO MIURA† and ISAMU SHIOYA††

In this investigation, we propose a new mechanism of Web Mining to extract events from a collection of Web Pages. Here we examine Web pages and obtain valid timestamp, and detect events by means of clustering. In this work, we apply TDT approach to Web environment and show how well our approach works to temporal Web pages by some experiments.

1. 動機と背景

近年のWebページの総量は莫大なものであり、日を追うごとに驚異的なスピードで増え続けている。この情報洪水の状況で、利用者はWebページ集合が何を表しているか理解することが難しくなる一方である。Webページ集合の表している内容について、いつ何が起こったのかを利用者が知っている場合も知らない場合もある。このためWebページ集合の内容を素早く容易に把握するための研究が近年注目を浴びている[2),9),10)]。

本稿ではWebページ集合からの自動的な事象抽出手法を提案する。現在、Google、Yahoo!等の検索エンジンを使えば、利用者は適切な検索語を与えることでいくつかのトピックを得ることができる。利用者にとって望ましい情報を見つけるのを手助けする為に、多くの検索エンジンは3億から30億と言われる巨大なURLデータベースを構築している。この巨大なデータベースを用いた検索により情報重複の問題を軽減させることができる。しかしながら、新たに検索結果のリストが長くなってしまうという問題が発生する。利用者は、得られた検索結果をブラウズし有益なWebページを探すのだが、多くの場合、途中で断念してしまう。実際、ほとんどの場合利用者は、最初の10又は20ページの中から有益なWebページを探し出すと言われており、この問題は深刻である。言い換えると、ページのランキングだけで出力されるべきである。現在では、参照の数、ハブとリンクなどのオーソリティ値、個人の好みなどの統計的な値を用いる手法などいくつかの手法が提案されている[5)]。

しかし、これらの手法はトピックを得るのに適した手法ではない。リストが示す内容を一見しただけで理解するのは困難であり、どれほどうまく並べられても、どのような事象が起きているかを理解することは難しい。解決法の1つとしては、ページを意味的にグループ化することが考えられる[4)]。検索したページをクラスタに分類し情報を要約できたならば、検索結果をより効果的に容易に吟味することができ、利用者の負担も軽減されると考えられる。

更に、ページの有効時間を類推することができれば、内容を事象ごとに理解することができ、Webページから時間軸上で自動的に事象を抽出することも可能になる。

この一連のアプローチを *Topic Detection and Tracking* (TDT) と呼ぶ[2),8)]。TDT研究プロジェクトでは、時間軸上で自動的にニュースストリームから

† 法政大学 工学研究科 電気工学専攻 〒184-8584 東京都小金井市梶野町 3-7-2
Dept.of Elect.& Elect. Engr., HOSEI University 3-7-2, KajinoCho, Koganei, Tokyo, 184-8584 Japan
†† 産能大学 経営情報学部 〒259-1197 神奈川県伊勢原市上粕屋 1573
Department of Management and Information Science, SANNO University 1573, Kamikasuya, Isehara city, Kanagawa 259-1197 Japan

トピックの意味の構造を抽出することを目的とした議論がされている。

ここでは、全ての情報が明示的もしくは暗黙のうちに時間の情報を持つと考えられる。時間の情報なくしては成り立たず、ここで扱うWebページに事象を検出するためにタイムスタンプを推定することは必要不可欠である。Webページに信頼性のあるタイムスタンプを類推できれば、事象又は傾向をより簡単にの検出することができる。

本稿では、URLの検索結果の並びを無視し、各Webページのタイムスタンプの類推を行う。通常，文書は，その内容に関する時間、すなわち**有効時間** (valid time) に従って理解されるが、必ずしも文書の内容時間が文書の作成・修正された時間、すなわち**作成時間** (creation time) や**トランザクション時間** (transaction time) と一致するものではない。

有効時間を類推しWebページを時間軸にしたがってクラスタ化すれば、個々のクラスタは意味のある事象に対応すると考えられる。本稿の、基本的な考えは、TDTのようにWebページをクラスタ化する観点から、適当な方法で事象を抽出することである。さらに、本稿では、個々のクラスタの意味解釈を自動的に与えるため、KeyGraphに基づく手法を用いる。

本稿では、2章でどのようにWebページから有効時間の抽出を論じる。3章で事象の抽出方法、その有効性を論じ、Googleを使い実験的な結果を論じる。4章で得られた事象を解釈するキーワードの決定の方法を論じ、5章で実験結果の考察を行い、6章で結論とする。

2. 有効時間の推定

本研究では文章の文法を解析することなくWebページから有効時間を類推する方法を提案する。ここでは3種類の時間を考慮する。(a) コンテンツに明示的に現れている内容時間 (CT)、(b) 作成時間 (UT)、(c) 更新時間 (TT) である。(b) はURLの一部に含まれており、(c) は受信ヘッダに明記してある。有効時間とは、Webページが示そうとしている時間を意味する。「松井稼頭央が2003/12/09にNew York Metsに入団」という文を例にとると、内容時間はWebページの文章に明示的に出現しているタイムスタンプである。この場合「2003/12/09」である。作成時間はWebページが生成された時間を言う。ここでは「2003/12/10」と仮定する。このとき、Webページの著者は原稿にしたがってページを生成するので内容時間と作成時間は必ずしも等しくならない。更新時間は、Webページが格納された時間又は、最後に更新された時間を言う。これを「2003/12/11」と仮定する。各Webページを解析し、内容時間、作成時間、更新時間を全て抽出しどれが有効時間により近いかを調査する。内容時間はそれぞれの文章の最初の文の前に現れるもので、「Jan 04, 2004」又は「January 3, 2004」のようなパターンのものを抽出する。複数の内容時間が抽出できる場合は、すべてを抽出する。次の例が示すように、経験的に作成時間はURLの一部として現れる。

`http://dsc.discovery.com/news/afp/20040105/marspix.html`

`http://www.cbsnews.com/stories/2004/01/04/tech/main591195.shtm`

最初のURLは2003/01/05の作成時間を含んでおり、次も同様に2004/01/04の作成時間を含んでいる。しばしば、URLは作成時間 (UT) を含む。更新時間については、受信ヘッダファイルに`"Last-Modified: Tue, 19 Aug 2003 06:10:54 GMT"` のようなLast-Modifiedのヘッダーが含まれる場合、そのWebページが「2003/08/19/06:10:54」で格納されたか、あるいは最後に更新されたことを意味する。だが、すべてのWebページのCT、UT、TT又はVTを必ず含むわけではない。有効時間を類推する為に、抽出したCT、UT、TTのどれがVTに近いか調査する。その為に、テキストコレクションを取得し、手作業によりそれぞれのページが有効時間を調べる。この時、全てのWebページが常にCT、UT、TT、VTを含むわけではないことに注意する。。例として、「under construction」が有効時間を持たなくても、UT又はTTを持つ場合がある。このとき、nullを用いて表す。一方、複数のCTを抽出できるWebページでは各CTのページのタプルを生成する、すなわちどのページも1つ以上のCTを持つことになる。

まず最初に、テストページpの集合Tを取得する：

$$T = \{< p_i, VT(p_i), ET(p_i), CT(p_i), UT(p_i), TT(p_i) > | i = 1, 2, ...\}$$

$T_P = \{p| < p, .. > \in T\}$ と定義する。

$p \in T_p$ が与えられたとき、ET(p) を推定するために、V, P_C, P_U, P_T, P_{CU}, P_{UC}, P_{CT}, P_{TC}, P_{UT}, P_{TU}, P_{CTU}, P_{CUT}, P_{TCU}, P_{TUC}, P_{UCT}, P_{UTC} : を次のように定義する：

$$
\begin{aligned}
V &= \{< p, VT(p) > | p \in T_P, VT(p) \neq null\} \\
P_C &= \{< p, CT(p) > | p \in T_P, CT(p) \neq null\} \\
&\cdots \\
P_{CT} &= P_C \cup \{< p, TT(p) > | p \in T_P, CT(p) \\
&\quad = null, TT(p) \neq null\} \\
&\cdots \\
P_{CTU} &= P_{CT} \cup \{< p, UT(p) > | p \in T_P, CT(p) \\
&\quad = null, TT(p) = null, UT(p) \neq null\} \\
&\cdots
\end{aligned}
$$

ここで V は全ての可能な答を意味する。他の定義はどのよう推定時間 (ET) を得るかを示す。例えば、P_{CU} は CT が null でない限り内容時間とし、$CT(p)$ が null だが $UT(p)$ が null でないときは $UT(p)$ を内容時間として類推する。この意味で、P_{CU} は内容時間の類推方法を示し、これを CU と示す。

Ans(答)、Rec(再現率) と Pre(適合率) を次のように定義する：

$$Ans(P) = \{< p, t > \in P \mid t = VT(p), t \neq null\}$$
$$Rec = |Ans(P)|/|V|$$
$$Pre = |Ans(P)|/|P|$$

Rec は、どれだけの答を $Ans(p)$ がカバーできたか、Pre はどれだけ正解を $Ans(p)$ が含んでいたかを示す。本稿では以下の式で示される F 値を用いる:

$$F = 2 \times Rec \times Pre/(Rec + Pre)$$

すべての組み合わせで F 値を算出し、実験的に最大の F 値のもの選択する。これを決定すれば、Web ページから類推時間を得る方法を求めたことになる。

本実験では、Google で `Kazuo matsui` を検索し、Top300 ページを得た。そこからリンク切れ、Weblog 以外の 235 の URL を対象とし、これらのページに対して手動でタイムスタンプを決定し 211 ページのタイムスタンプを得た。そして、各 Web ページから内容時間 CT、作成時間 BT、更新時間 TT の抽出を行った。スキーマは、内容時間の類推方法を示し、ExpTime は null ではない時間を持つページの数であり Ans はスキーマごとの答の数を示す。

この結果からわかるように、UC が Pre 値が最も高い値であるが、F 値が最大となるスキーマは CU である。CU は、多くの答えをカバーしている事を意味する Recall の値が最も良い。一方 CUT の F 値は TT の Pre 値が非常に悪いために小さくなっている。

3. 事象の抽出

Web ページ集合から有効時間 (VT) を推定したが、次に事象抽出を行う。この章では、類推した VT を用いて Web ページの事象 (Event) を抽出する方法を提案する。

TDT の分野において、時間軸におけるクラスタ化が効果的であるとはよく知られている[2]。すなわち、事象はしばしば時制クラスタに対応する。以下では Web ページ集合を検索エンジンに検索語を与え、その結果を得ており、Web ページ集合は 1 つのトピックについて論じられていると考えられる。

ここでは時間軸でクラスタ化することの正当性を、`Kazuo Matsui` のページを用いて示す。235 の Web ページから最も F 値の高かった CU スキーマにより取得した 177 のページに対し、K-means アルゴリズムを利用してクラスタ化を行う。Page はクラスタ内のページの数を示し、CT は、内容時間のページの数、UT は、作成時間のページの数を示す。以下はその結果である。

Group	Time Interval	Pages	CT	UT
Group0	1975/10/23 - - 1975/10/23	5	5	0
Group1	1995/06/20 - - 1997/11/18	4	4	0
Group2	2000/06/27 - - 2001/11/04	4	2	2
Group3	2002/03/16 - - 2003/01/01	5	5	0
Group4	2003/06/29 - - 2003/12/20	93	87	6
Group5	2003/12/27 - - 2004/01/26	24	22	2
Group6	2004/01/31 - - 2004/02/17	17	15	2
Group7	2004/02/19 - - 2004/03/06	25	24	1
(total)		177	164	13

図 1 に結果を示す。結果として、8 つのクラスタを得た。半分のクラスタは、クラスタの構成要素が 5 以下であり非常に小さいので無視する。残る各 4 つのクラスタを解釈する為に、`Kazuo Matsui` を含む文を選

Scheme	ExpTime	Ans	Pre	Rec	F
C	164	127	77.4	60.2	67.7
U	52	42	80.8	19.9	31.9
T	68	2	2.9	0.9	1.4
CU	177	133	75.1	63.0	68.6
CUT	213	133	62.4	63.0	62.7
CT	202	127	62.0	63.0	62.5
CTU	213	132	62.0	63.0	62.5
UC	169	130	76.9	61.6	68.4
UCT	205	130	63.4	61.6	62.5
UT	113	43	38.1	20.4	26.6
UTC	184	105	57.1	49.8	53.2
TC	192	102	53.1	48.3	50.6
TCU	203	105	51.7	49.3	50.5
TU	113	38	33.6	18.0	23.5
TUC	184	93	50.5	44.1	47.1

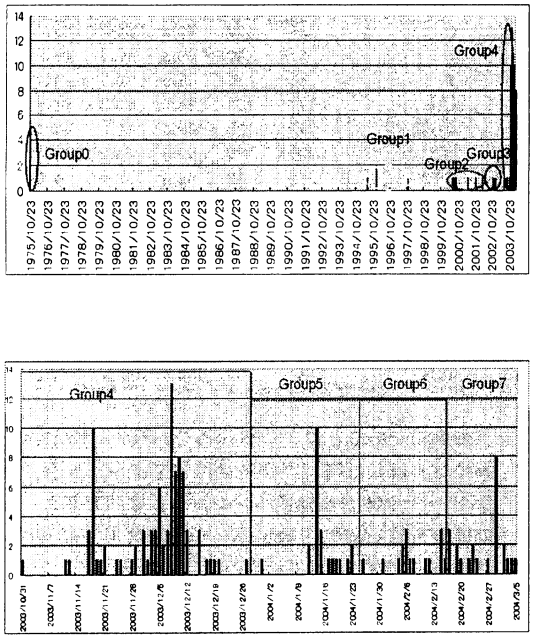


図 1 Clustering Kazuo Matsui pages

択して、頻繁な語、特徴的な語を手作業でを取り出した。93 ページからなる Group4 のうちいくつかの文章を例として示す。

```
http://www.bayarea.com/mld/cctimes/sports/7289763.htm
  Seibu Lions shortstop Kazuo Matsui wants to play in the major
  leagues, the seven-time Japanese League All-Star said Monday.
http://www.boston.com/sports/baseball/articles/2003/12/09/kaz_
matsui_signs_on_with_the_mets/
  Kaz Matsui signs on with the Mets
http://www.taipeitimes.com/News/sport/archives/2003/12/12/
2003079339/print
  NY Mayor welcomes Matsui No. 2
....
```

これらのクラスタの文章より、このクラスタを **Mets welcome sign major league** と我々が解釈した。全てのクラスタの解釈を下に示す。

(Group4: 2003/6/29 - 2003/12/20)
Mets welcome sign major league
(Group5: 2003/12/27 - 2004/1/26)
ready challenge
(Group6: 2004/1/31 - 2004/2/1 spring opening exhibition game
(Group7: 2004/2/19 - 2004/3/6)
injure finger champ

すべてのクラスタの解釈は **Kazuo Matsui** に関したいくつかの事象で適当なクラスタであり妥当であると言える。すなわち、Web ページから事象を抽出するのに提案手法は有用である。時制的な側面を持つ Web ページにも TDT と同じ傾向を持つことを示している。

4. 事象の解釈

Web からのクラスタを自動解釈する試みを提案する。本稿では、各クラスタから重要な語句を抽出し、これをラベルとする方法をとるが、ここでは KeyGraph の考え方を用いる。

KeyGraph とは、文章中に出現する単語の出現頻度と共起関係から文章の主張点を把握し、キーワードを抽出する手法である[7]。KeyGraph では、文章中に頻繁に出現する言葉は文章が書かれる上での前提条件、つまり基本的な概念であり「土台」と呼ぶ、更に「土台」によって支えられている語が文章の「主張」(筆者の主張) であると見なす。KeyGraph の生成は、いくつかのステップからなる。まず、文章から不要語を取り除き、上位定数個の頻出単語間の共起度を計算し語同士の間に定数個の共起リンクを張る。次に、共起度の薄いリンクを切断して土台となる語のグループを作る。土台の語グループと、不要語を取り除いた文章中に出現するすべての語の共起度を算出し、値の高い定数個の語 (主張) の共起リンクを残す。更に、共起リンクの設定された、土台となる語と主張語の共起度を計算し、共起リンクに値を与え共起リンクの和をとる。これを土台と主張を結びつける重要語とみなし上位語を選定する。

本稿では、この KeyGraph により得られる主張語を時制クラスタの解釈に用いる。実際、時間軸でクラスタリングされた Web ページは相互に類似性が高く、得られた主張語集合には極端な差異は生じない。ただ、時間軸に沿って変化している様をとらえるため、1 つ前のクラスタの主張語集合と比較し、その差分で時制クラスタを解釈する。このとき、時間軸で一番古いクラスタは差分計算ができないので、以下では、主張語として得られた全ての語の、上位 7 パーセントを差分対象に用いる。

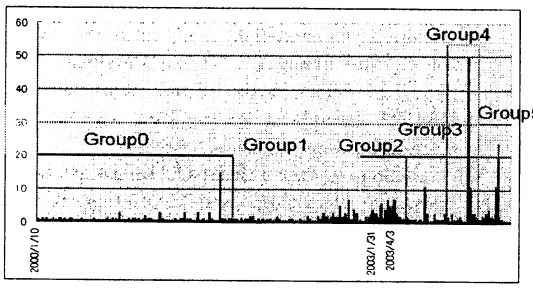
5. 実　　験

提案手法の有用性を示すために、Google によって得られる 1000 ページの Web ページについて実験的な結果を論じる。先に述べた様に、最初に URL のリストを取得し、提案した手法 CU にしたがって VT の類推を行う。

5.1 時制クラスタの抽出

本実験では、検索語 hussein という条件の下に Google より 1000 個の URL リストを取得し、リンク切れ、Weblog、時間情報のないページを取り除いた結果、669 のページを得た。この 669 ページを時間軸によりクラスタ化すると、図 2 で示すような 6 つのクラスタを得た。

GroupID	Pages	ContentTime	URL Time
Group0	82	75	7
Group1	101	79	22
Group2	162	129	33
Group3	57	51	6
Group4	182	156	26
Group5	85	80	5
Total	669	570	99



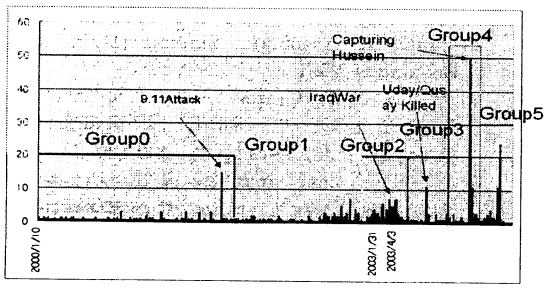


図 2 Clustering Hussein pages

2001/12/15 と 2002/11/20 の間の 101 ページの Group1 の文の例を示す。これら、すべてがサダム・フセインのいくつかの様相について述べている。

```
The U.S. Must Strike at Saddam Hussein
Bush planning to topple Hussein
Saddam Hussein to be overthrown by the opposition
Opposing Saddam Hussein
[Hussein Ibish:] U.S. Arabs' Firebrand
IRAQ: CRIMES AGAINST HUMANITY
Leaders as Executioners
How The US Armed Saddam Hussein With Chemical Weapons
Peasant-born Saddam relentlessly pursued prestige, power
For decades, Iraqi leader was both omnipresent, elusive
Hundreds Show Up For Anti-Hussein Rally
Bin Laden Linked To Saddam Hussein
....
```

これらのクラスタに対して次のように解釈した。

(Group0: 2000/01/10 - 2001/12/18) Attacks on World Trade Center and Pentagon
(Group1: 2001/12/28 - 2002/11/27) About Saddam Hussein
(Group2: 2002/12/02 - 2003/05/14) Start War
(Group3: 2003/05/19 - 2003/10/03) Uday and Qusay were killed in a battle with U.S.
(Group4: 2003/10/08 - 2004/01/22) Saddam Hussein captured
(Group5: 2004/01/26 - 2004/03/22) After Getting Hussein

これらの解釈は非常にクラスタに対応したものであると言える。実際、図で示されるように、特有の問題は適切なクラスタで発生している。

5.2 時制クラスタの解釈

次に、上述クラスタを KeyGraph 手法を用いて主張語を取り出し、差分を抽出する。

クラスタ	主張語
0	31
1	50
2	54
3	40
4	63
5	34

次はクラスタ 1 の (クラスタ 0 との) 差分である。

weapon, militari, iran, 2002, inspector, intern, bush, document, famili, russian, nuclear, washington, threat, 2003, offici, 16, kamel, christiansciencmoni, tore, claim, control, defect, march, missil, opposit, terrorist, plan, terror, senat, agreement

これらステミングされた状態であるので、そのままでは理解しにくいが、さらに得られた主張語集合は、辞書や背景知識などを用いて抽象化・集約化されて統合できる☆。ここでは、これを次のように人手で要約する:

武装:weapon, military, plan
国際:russian, iran, internaltional,
アメリカ国内:senat, bush, tore, claim, control. defect washington
UnitedNations:document, inspector, agreement, terro, terrosist nuclear, missile, opposite, threat
報道:ChristianScienceMonitor

☆ たとえば Wordnet などの辞書を活用すればよい。

イラク:famili, kamel

これらの内容は、上述の人間による解釈 (About Saddam Hussein) を相当程度精密に記述したものである。

同様に、クラスタ 2 (Start War) は次のような主張語と対応している。

武装:enemy, capture, attempt, army, defense, aggressive

国際:world

アメリカ国内:leader, nation

イラク国内:author, coalit, Kurd, Bin-Laden, Amicu, Dictator party

報道:report, talk, live, fact

UnitedNations:WeaponMassDestruction, Answer

クラスタ 3(Uday and Qusay) も同様に次のような主張語と対応する。

武装:recruit, military,oper, troop

ウダイとクサイ:July, Husseins, son, udai,qusai

イラク体制:bremer, power, intelligence, intelligentserv, mukhabarat, secure,

クラスタ 4 (Saddam Captured) では特に報道分野の語が現れる。

武装:soldier, attempt

国際:arab, world, countries, intern

アメリカ国内:bush,polit, polici

UnitedNations:weapon, document

報道:video, article, report, copyright, ChristianScienceMonitor, site, work

フセイン:captur, family, sunday, death, trial, hole, crime, tikrit

イラク体制:administr, govern, leader, nation, coalit, regim

クラスタ 5(after getting Saddam) ではその後の状況変化を捉えた語が現れる。

往来:visit, com

支援・体制:redcross, author , ICRC

UnitedNations:ICRC, evid

これらから判断し、得られた語は、予め与えた解釈を精緻に述べるものであり、直感的に捕らえやすいものとなっている。以上から、提案手法が我々の知識で有効性が示せたと言える。

6. 結　　論

本稿では、検索語を与え検索エンジンから Web ページ集合を取得し、時制 Web ページ集合から事象の抽出を行う方法と KeyGraph を用いてクラスタの自動解釈を行う方法を提案した。

最初に、Web ページの有効時間の類推を行い、小さなテストページ集合で予備実験を行い、経験的に類推方法 P_{CU} を採用した。次に、K-means アルゴリズムによりクラスタを作り、KeyGraph の方法に基づいてそれらの解釈を行った。実験に基づく結果は、時制 Web ページで手法が有効であることを示し、時制 Web ページから正確で適切に事象を抽出できることを意味している。有効時間を類推できれば、事象の検出と追跡も容易なると考えることができる。

参 考 文 献

1) Alexandrin Popescul,Lyle H. Ungar.: Automatic Labeling of Document Clusters,unpublished

2) Allan, J., Carbonell,J., Doddington, G., Yamron,J. and Yang,Y.: Topic Detection and Tracking Pilot Study: Final Report, proc. DARPA Broadcast News Transcription and Understanding Workshop (1998)

3) Grossman,D. and Frieder,O.: Information Retrieval – Algorithms and Heuristics, Kluwer Academic Press, 1998

4) Jain, A.K., Murty, M.N. et al.: Data Clustering, *ACM Comp.Surveys* 31-3, 1999, pp.264-323

5) Kleinberg, J.M. : Authoritative Sources in a Hyperlinked Environment, JACM 46-5, 1999

6) Mani, I.: Automatic Summarization, John Benjamins, 2001

7) 大沢幸生：KeyGraph　一語の共起グラフの分割統合によるキーワード検出、電子情報通信学会論文誌 D-I、J82-D-I2,pp.391-400,1999

8) NIST (National Institute of Standrads and Technology): www.nist.gov/speech/tests/tdt/

9) Radev, D. and Fan, W. : Automatic summarization of search engine hit lists, proc ACL'2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, 2000, Hong Kong

10) Yang, Y., Pierce, T. and Carbonell,J.: A Study on Retrospective and On-Line Event Detection, proc. SIGIR-98, ACM Intn'l Conf. on Research and Development in Information Retrieval, 1998