6-2007

Selecting Labels for News Document Clusters

Krishnaprasad Thirunarayan
Wright State University - Main Campus, t.k.prasad@wright.edu

Trivikram Immaneni

Mastan Vali Shaik

Selecting Labels for News Document Clusters

Krishnaprasad Thirunarayan, Trivikram Immaneni and Mastan Vali Shaik

Metadata and Languages Laboratory
Department of Computer Science and Engineering
Wright State University, Dayton, Ohio-45435, USA.
{t.k.prasad, immaneni.2, shaik.7}@wright.edu
http://www.cs.wright.edu/~tkprasad

Abstract. This work deals with determination of meaningful and terse cluster labels for News document clusters. We analyze a number of alternatives for selecting headlines and/or sentences of document in a document cluster (obtained as a result of an entity-event-duration query), and formalize an approach to extracting a short phrase from well-supported headlines/sentences of the cluster that can serve as the cluster label. Our technique maps a sentence into a set of significant stems to approximate its semantics, for comparison. Eventually a cluster label is extracted from a selected headline/sentence as a contiguous sequence of words, resuscitating word sequencing information lost in the formalization of semantic equivalence.

1 Introduction

A scalable approach to processing large document datasets (such as Medline, News documents, etc.) can be obtained by indexing and classifying the documents by stamping each document with metadata terms from a well-defined ontology that reflects and abstracts the document's content, and then manipulating only the metadata terms in lieu of the document content. For instance, the UMLS[1] terms (metadata) can be used to construct and label a related set of Medline documents involving certain genes, diseases, or organs [14], and the SmartIndex tags (weighted metadata) can be used to construct and label a related set of News documents involving certain entites or events [12]. Metadata-based cluster labels can be significantly improved to better indicate the content of the document clusters obtained in response to entity-event search queries, by generating labels that are grounded in and extracted from the document text of the document clusters. This paper presents a simple technique to construct and select good cluster labels in the context of News documents obtained in response to search queries involving entities and events.

As an illustration, consider sentence fragments and headlines in recent News about the entity "Nokia" and the event "Mergers and Acquisitons".

– Nokia acquires Intellisync.

[1] Unified Medical Language System

- Intellisync to be acquired by Nokia.
- Nokia's (NOK) acquisition of Intellisync (SYNC) will not change the overall picture for company, says Greger Johansson at Redeye in Stockholm.
- The acquisition of Intellisync supports Nokia's goal to be the leader in enterprise mobility and enhances the ability of its customers to connect devices to data sources, applications and networks.
- Nokia and Intellisync have signed a definitive agreement for Nokia to acquire Intellisync.
- Nokia's Intellisync buy.
- Nokia's purchase of Intellisync.

Besides a reference to the explicitly searched entity (that is, "Nokia") and the event (that is, "acquires", "buy" etc.), the cluster label should contain other relevant information (such as "Intellisync") about the queried subjects (analogous to answer extraction), to provide a concise highlight of the document collection. For the above example, electing "Nokia acquires Intellisync" seems reasonable for the following reasons:

- It is *sound*, containing "Nokia", and a reference to "Mergers and Acquisions" via "acquires".
- It is *complete*, containing additional relevant information "Intellisync".
- It is *well-supported*, with majority of the document fragments providing supporting evidence for it.

The issue of selecting cluster labels natually arises in the context of construction of timelines of trends (e.g., Google Trends [11] response for "Chemistry vs Physics"), implementation of entity-event-duration timelines with call-out labels (e.g., Microsoft and Mergers & Acquisitions in the Year 2005) from the News Document dataset, etc.

In summary, we address the issue of formalizing sentence fragments of News documents that abstracts the meaning of sentences adequately and is lightweight so as to be scalable. Section 2 formalizes the selection of "good" cluster labels in two steps: Section 2.1 motivates and specifies the selection of a promising sentence and analyzes various other alternatives to justify the superiority of the chosen criteria. Section 2.2 explains how to delimit a concise and comprehensible label from the chosen sentence. Section 3 considers the restricted situation when only document headlines are available but not the document contents, for proprietary reasons. Section 4 discusses the implementation of the Timeline application in News documents context. Section 5 briefly reviews some of the recent related work. Section 6 concludes with suggestions for future work.

2 Construction and Election of Cluster Labels

We make the pragmatic assumption that the documents in a News document cluster contain sentences and/or headlines that can yield cluster labels, and propose an approach to selecting cluster labels by extracting a content phrase from a chosen high-scoring sentence or headline containing the queried subjects (entities and events). We assume that sentences include the headline.

2.1 Sentence Selection from Metadata Screened Document Cluster

Problem Statement: Consider a cluster of documents $CD = \{D_1, D_2, ..., D_m\}$ for an entity EN and an event EV. Extract, from the cluster documents, a *well-supported* sentence that contains phrasal references to EN and EV.

Informally, a well-supported sentence is obtained by maximizing the number of documents that support the sentence, by maximizing the degree of overlap with a sentence in each document, and by minimizing its length, subject to the constraint that the sentence contains phrasal references to EN and EV. The rest of this section assumes the existence of such sentences and formalizes the notion of well-supportedness. (In other words, this paper takes a shallow, scalable approach to extracting a "good" label, as opposed to understanding the content of the document cluster and synthesizing a label from its meaning.)

Proposed Approach:

1. Let $sen(D_i)$ refer to the set of sentences in document D_i that each contain phrasal references to the entity EN and the event EV. We call them significant sentences of the document D_i.
   The phrases corresponding to an entity or an event can be obtained from the domain knowledge used to stamp the documents with metadata terms. This part of the domain knowledge is encapsulated as Metathesaurus in UMLS or as Concept Definitions in the context of News documents. It includes synonyms, acronyms, and other equivalent usages for a metadata term. A mature indexing and search engine can be used to determine if an entity phrase and an event phrase co-occur in a sentence (or a paragraph) of a document. (Otherwise, such an engine can be built using open source APIs such as Apache Lucene [13] by materializing sentence/paragraph separators. We skip the implementation details here because they are peripheral to the main theme of this paper.) In the case of News documents, applications can be built to determine and extract metadata terms from each sentence of a document using special purpose indexing APIs. Note also that the co-occurrence within a sentence (or within a paragraph) is a better yardstick of semantic relationship than just the incidental positional proximity of an entity phrase and an event phrase in the document text that may be the result of co-occurrence in two neighboring sentences or in two neighboring paragraphs (worsened for compound News documents containing multiple stories).
2. To each sentence s, we associate an abstraction (meaning), $M(s)$, a set of strings that best approximates the semantics of s. For specificity, let $M(s)$ be the set of stems obtained from s as follows: collect all the words in s into a set, eliminate the stop words, and then stem each word. The rationale is that this normalization better captures the semantics (that can be used to check for semantic similarity between sentences through purely syntactic manipulation). Observe that:
   - This removes overly discriminating word sequencing information from a sentence such as due to active/passive voice changes.

- It is not doomed by the well-known problems associated with the bag-of-words model of documents because the focus is only on sentences, and not the entire document.
- The word sequencing information will be resuscitated in Section 2.2 while generating the final terse cluster labels.

Other alternatives that we regard as inferior to $M(s)$ are:

(a) $M_1(s)$ is the (non-contiguous) *sequence* of words of s obtained by eliminating the stop words. This normalization is inadequate because it is overly discriminating. For example, consider "The merger of America Online and Time Warner will create the world's largest media company" vs "The merger of Time Warner and America Online will create the world's largest media company". These two sentences do not match because Time Warner and America Online have been swapped.

(b) $M_2(s)$ is the *set* of words from s obtained by eliminating the stop words. This normalization is an improvement over $M_1(s)$ but it has its limitations when we consider semantics-preserving voice changes (active to passive, and, passive to active). For example, consider "Nokia acquires Intellisync." vs "Intellisync acquired by Nokia." vs "Nokia's acquisition of Intellisync." $M(s)$ is an improvement over $M_2(s)$ because stemming comes to rescue, treating "acquires", "acquired" and "acquisition" as equivalent. Recall that stemming reduces different forms of a word to the same root, and the risk of two semantically different words getting reduced to the same stem (using scalable Porter stemming algorithm) is minimal, in our limited context.

(c) $M_3(s)$ can be defined in such a way that two words are treated as equivalent if they have the same stems or have been asserted so via the codified domain knowledge (that is, in the Metathesaurus or through the concept definition). For example, consider "Nokia acquires Intellisync." vs "Nokia buys Intellisync." vs "NOK purchases SYNC." If the domain knowledge implies that "acquires", "buys", and "purchases" are synonymous in the "Mergers and Acquisition" context, all the three phrases are equivalent.

Note that $M_1(s)$ refines $M_2(s)$, $M_2(s)$ refines $M(s)$, and $M(s)$ refines $M_3(s)$.

3. To compute the support a document D_j accords to a sentence $s \in sen(D_i)$, we use the following scoring strategy[2]:

$$score(s, D_j) = \frac{MAX_{t \in sen(D_j)} |M(s) \cap M(t)|}{|M(s)|}$$

and for cumulative support score for sentence s due to the entire cluster

$$score(s) = \Sigma_{j=1}^{m} score(s, D_j)$$

Observe that:

- The support that a document D_j accords to a sentence s is proportional to the maximum number of overlapping stems in a D_j-sentence, scaled by the number of significant stems in s. As such, each document D_j can contribute at most 1 to the score for s.

[2] $|...|$ is the set cardinality function.

- The appearance of significant stems of s in t is a plus. Also, $M(s) \subseteq M(t) \Rightarrow score(s) \geq score(t)$.
- If every document contains the same two significant sentences, then both sentences will garner equal score irrespective of the document content or length. (In the context of "on-topic" News, this does not lead us astray.)
- However, if there is one document D that contains a sentence s with smaller number of significant stems than a sentence t, then the contribution from D to the overall score of s will be higher than that for t, due to the scaling with respect to the number of significant stems. That is, $M(s) \subset M(t) \Rightarrow score(s) > score(t)$.

Reconsider the example in Section 1, reproduced below for convenience.

S0: Nokia acquires Intellisync.
S1: Intellisync to be acquired by Nokia.
S2: Nokia's (NOK) acquisition of Intellisync (SYNC) will not change the overall picture for company, says Greger Johansson at Redeye in Stockholm.
S3: The acquisition of Intellisync supports Nokia's goal to be the leader in enterprise mobility and enhances the ability of its customers to connect devices to data sources, applications and networks.
S4: Nokia and Intellisync have signed a definitive agreement for Nokia to acquire Intellisync.
S5: Nokia's Intellisync buy.
S6: Nokia's purchase of Intellisync.

According to our scoring criteria, the sentences [S0:] and [S1:] are treated as equivalent. They are supported by [S2:], [S3:] and [S4:], and to a much lesser degree by [S5:] and [S6:]. Furthermore, [S0:] and [S1:] score higher than [S2:], [S3:] and [S4:] on the basis of their "scaled" length. All this seems reasonable because we prefer a sentence that has strong "verbatim" support from document sentences of the cluster.

4. A well-supported sentence s for cluster label for the cluster of documents CD is the one that has the maximum cumulative support score.

$$candidate?(s) = \forall t \in \cup_{j=1}^{m} sen(D_j) : score(s) \geq score(t)$$

$$well\_supported\_sentences(CD) = \{\ s \in \cup_{j=1}^{m} sen(D_j)\ \mid\ candidate?(s)\ \}$$

2.2 Phrase Selection for Cluster Label

A *candidate cluster label* is the shortest sequence of words (in terms of string length) of a sentence in a document that contains a phrasal reference to the entity EN, the event EV, and significant words that appear in all well-supported sentences.

$$common\_stems(CD) \ = \ \cap \{M(s)\ \mid\ s \in well\_supported\_sentences(CD)\}$$

$$label\_pool(CD) \ = \ \{ss\ \mid\ \exists s \in well\_supported\_sentences(CD) \wedge substring(ss, s)$$

$$\wedge\ [\ \forall t \in common\_stems(CD):\ substring(t, ss)\ ]$$
$$\wedge\ preceded\_and\_followed\_by\_delimiter(ss)$$
$$\wedge\ contains\_entity\_event\_reference(EN, EV, ss)\}$$

Note that

- *common_stems*(_) check has been incorporated to extract meaningful additional information to highlight,
- *preceded_and_followed_by_delimiter*(_) check (that determines whether the preceding and succeeding character is a blank or a punctuation mark) has been incorporated to generate a sequence of words as opposed to some substring, and
- *contains_entity_event_reference*(_) check has been incorporated to ensure that the label contains some concrete reference to queried entity and event. This check can be as simple as a verbatim match to as complex as requiring alias resolution (for example, involving acronyms, coreferences, etc).

$$candidate\_cluster\_labels(CD) \ = \ \{p \in label\_pool(CD)\ \mid\ \forall q \in label\_pool(CD):\ |p| \leq |q|\}$$

A *cluster label* can be any one of the candidate cluster labels.

Observe that the cluster label is required to be a *contiguous* sequence of words from a well-supported sentence of a document in the cluster, and that it is not unique in general. For example, if *label_pool*(CD) contains "Nokia acquires Intellisync", "Intellisync to be acquired by Nokia", "Nokia's (NOK) acquisition of Intellisync (SYNC)", "acquisition of Intellisync supports Nokia's", and "Nokia to acquire Intellisync", then "Nokia acquires Intellisync" is the chosen cluster label. For the dataset containing {S0, S5, S6}, each of S0, S5 and S6 is in *label_pool*(CD) but only S5 can be the cluster label.

Other alternative is to define *cluster label* as a shortest label (in terms of number of words) among those in the cluster label pool. For the first example, "Nokia acquires Intellisync" again wins the cluster label competition. For the dataset containing {S0, S5, S6}, both "Nokia acquires Intellisync" and "Nokia's Intellisync buy" are equally acceptable as cluster labels, while "Nokia's purchase of Intellisync" is rejected (on the feeble grounds that it contains the stop word "of"). For the dataset containing {S2, S3, S4}, "Nokia to acquire Intellisync" is the selected cluster label.

3 Selection of Cluster Labels from Headlines Alone

For proprietary reasons, only the metadata associated with a News document that includes the headline may be available, instead of the entire document or the APIs for extracting metadata from sentences/documents. Cluster labels can then be generated from headlines alone by adapting the above solution.

Problem Restatement: Consider a cluster of documents with headlines $\{D_1, D_2, ..., D_m\}$ for an entity EN and an event EV. Extract a *well-supported* headline from the documents of the cluster as follows.

Modified Approach:

1. For each $i \in 1 \ldots m$, let h_i refer to the headline of the document D_i.
2. Similarly to the approach discussed in Section 2.1, to each headline h, we associate an abstraction, $M(h)$, a set of strings that best approximates the semantics of h. *Define $M(h)$ as the set of stems obtained from h by collecting all the words in h into a set, eliminating the stop words, and then stemming each word.* This normalization removes overly discriminating word sequencing information from a headline, and generates word forms that are better amenable to syntactic manipulation for gleaning semantics.
3. To compute the support a document D_j accords to a headline h_i, we use the following scoring strategy:

   $$score(h_i, D_j) = \frac{|M(h_i) \cap M(h_j)|}{|M(h_i)|}$$

   and for the cumulative support score for headline h_i due to the entire cluster

   $$score(h_i) = \Sigma_{j=1}^{m} score(h_i, D_j)$$

   Observe that: $M(h_i) \subseteq M(h_j) \Rightarrow score(h_i) \geq score(h_j)$.
4. A well-supported headline h for cluster label is the one that has the maximum cumulative support score.

   $$well\_supported\_headlines(\cup_{j=1}^{m}\{D_j\}) = \{h \mid \forall j \in 1 \ldots m : \ score(h) \geq score(h_j)\}$$

   Any one of the well-supported headlines can be used as a cluster label. Note that in entity-event-duration timeline application, there is usually a unique candidate headline for a cluster of documents involving an entity and an event on a day because several news stories are correlated.

4 Application Context and Implementation Details

We have implemented an entity-event-duration timeline application in Java 5. The application pre-processes a year's worth of metadata-tagged News stories (provided in the form of XML documents) (150GB), indexing them for efficient access. The application takes an entity, an event, and a time-duration, and generates a timeline based on the number of News stories involving the entity and the event. As depicted in Figure 1, for each date, the GUI can pop-up a listbox showing the headlines of all the relevant documents and a generated cluster label. (It can also display the contents of a chosen News document.) The prototype works acceptably in practice, and we are investigating quantitative metrics to evaluate such systems in the absence of standard benchmarks or human analysts.

We now discuss several concrete examples to bring out the nature of the News documents and the behavior of our prototype. (The examples were chosen to be realistic as opposed to idealistic.) For example, on *April 12, 2005*, for the entity *Microsoft*, and for the event *Computer Operating Systems*, the generated headline cluster label is: *In next Windows release, Microsoft to use hardware for security*, based on the headlines:

Fig. 1. Entity-Event-Duration Timeline Application

Microsoft unveils more details of next Windows release
In next Windows release, Microsoft plans to use hardware to lock down security
In next Windows release, Microsoft to use hardware for security
Microsoft ships Windows for 64-bit computers
Microsoft Gives Details on Windows Release
New Windows Operates on 64-Bit Computers
Microsoft ships Windows for 64-bit computers
In next Windows release, Microsoft to use hardware for security
Microsoft unveils more details of next Windows release
Microsoft ships Windows for 64-bit computers
Microsoft unveils more details of next Windows release
In next Windows release, Microsoft will use hardware for security
Microsoft plans to use hardware to lock down security in Windows
Microsoft ships Windows for 64-bit computers
In next Windows release, Microsoft to use hardware for security
Gates shows off features of next-generation Windows system

Our criteria effectively chooses the most frequent "short" headline such as *In next Windows release, Microsoft to use hardware for security* (or *In next Windows release, Microsoft will use hardware for security*) as the cluster label, while

ignoring other headlines such as *Microsoft ships Windows for 64-bit computers*. (The majority criteria was chosen to eliminate "noise".) Several documents share a headline due to correlated News sources (such as Associated Press, Reuters, AFX News, etc.). Each such document can be viewed as providing an independent endorsement. Unfortunately, this approach can miss multiple headlines for different News stories that happen to have the same event and entity metadata tags, and occur on the same day. In fact, the "best" comprehensive headline for the above example is: *Microsoft unveils more details of next Windows release*. So, our approach can be further improved by clustering headlines on the basis of similarity, or ranking headlines on the basis of support and cutoff thresholds.

The other approach to cluster label generation elects a significant sentence from the cluster documents and clips it. For example, on *April 4, 2005*, for the entity *BHP Billiton*, and for the event *Takeovers*, the relevant document sentences from three separate News documents are:

1. Anglo-Australian resources giant BHP Billiton has been given the green light by Treasurer Peter Costello for its $9.2 billion takeover of Australian miner WMC Resources.
2. WMC had been the focus of a hostile takeover by Swiss-based Xstrata that had gained attention from the government backbench before BHP put in its bid.
3. Mr Costello, who had the ability to block the takeover or set impossible restrictions, only set two conditions on BHP and its proposal, both relating to uranium.
4. BHP chief executive Chip Goodyear welcomed the decision, saying the treasurer's conditions were acceptable and the company would abide by them.
5. BHP has offered $7.85 for each WMC share, with the takeover bid due to close at 7.30 pm (AEST) on May 6.

1. The federal government had raised no objection to the proposed takeover of WMC Resources by BHP Billiton, Treasurer Peter Costello said today.
2. In a statement, Mr Costello set two conditions for the proposed $9.2 billion takeover of WMC by BHP Billiton.

1. BHP Billiton chief executive Chip Goodyear welcomed the government's approval of the WMC bid.
2. The company said the conditions attached to the announcement by the Treasurer today were acceptable to BHP.

The well-supported sentence to summarize the cluster is: *In a statement, Mr Costello set two conditions for the proposed $9.2 billion takeover of WMC by BHP Billiton.*, yielding the cluster label: *takeover of WMC by BHP Billiton*.

Our cluster label generator can also implement $contains\_entity\_event\_reference(\_)$ using tagging APIs that looks for co-occurrence of the queried terms in a paragraph or in a sentence, as opposed to using existing document-level XML tags. (We do not have license to use proprietary concept definitions (associations between metadata terms and document phrases employed by the tagger) to develop

our own tagger.) To see the limitations of the current sentence-based approach, it is instructive to consider cluster labels generated from the 2005 News dataset given below. Note that we have intensionally excluded headlines to see how reliable document sentences are in yielding suitable cluster labels. If the headlines were included, they seem to dominate the cluster labels for obvious reasons.

Entity	Event	Date	Cluster Label
Toyota	Automotive Sales	June 3, 2005	Toyota Motor Corp. posted a 0.5 percent sales drop to 201,493 units
Google	Mergers & Acquisition	April 22, 2005	Google, (GOOG) the No. 1 search engine, said its first-quarter profit
Google	Internet & WWW	June 22, 2005	Google, which depends upon online
Google	Search Engine	June 22, 2005	Google to sell content through its search engine
Google	Online Advertising	June 22, 2005	Google will develop another source of revenue besides online advertising
Sprint	Mergers & Acquisition	June 2, 2005	Sprint Corp. shareholders are expected to vote in early July on the company's planned merger

In order to see the reliability difference between paragraph level vs sentence level co-occurrence for inferring associations, consider the document for entity *Microsoft* and event *Mergers & Acquisition* on *April 19, 2005* containing the fragment:

> ... Amazon already offers e-books and more than 1 million e-documents on its site, using downloadable software from *Microsoft Corp.* and Adobe Systems Inc. The *purchase* of Mobipocket will allow Amazon to use its own software to diversify product distribution methods, rather than relying on third-party providers. ...

The indexing metadata tags *Microsoft* and *Mergers & Acquisition* are associated with the phrases 'Microsoft' and 'purchase'. If document-level or paragraph-level co-occurrence of phrases is used for inferring associations, we get a false positive. As the phrases appear in successive sentences, sentence-level co-occurrence can improve reliability.

5 Related Work

Our entity-event-duration timeline application resembles Google Trends [11] which tries to determine relative interest in a topic on the basis of the number of searches on the topic (Search volume graph) and the number of News stories involving the topic (News reference volume graph). It also summarizes search query distribution in terms of their geographical location of origination (such as city, country, etc), language of the search query, etc. The spikes in search volume graph are further annotated by the headline of an automatically selected Google News story written near the time of that spike. Our entity-event timeline interface allows you to display all the News documents for the year 2005 that

carry the corresponding entity and event index terms (with scores higher than a programmable cutoff).

Our work on Timeline generation can be viewed as a means to cluster search results using temporal attribute which happens to be the News story creation date [1]. Our label generation work is related to Vivisimo's post-retrieval tagging that provides more meaningful labels than those found in general tagging vocabularies [16].

Several proposals use frequest-item sets to derive labels [2, 6]. Even though these approaches have a more general appeal, our approach provides more comprehensible and comprehensive labels in the context of News documents because it takes into account the semantics of the words using encoded domain knowledge, and the sequencing of the words by extracting from documents a concise phrase containing the "frequent items" to serve as the cluster label.

The techniques for cluster label generation described in [3, 4, 7–9] deal with the problem of abstracting sequencing information to improve precision and to rank labels in the general text documents context. Our approach addresses this issue by first using set-based abstraction to deal with semantic equivalence problem, and eventually restore word sequencing information to arrive at palatable labels in the more restrictive News documents query results set context.

The cluster description formats discussed in [10] are similar in spirit to our work but it deals with clusters of numerical records rather than text documents.

QCS information retrieval system [5] extracts documents relevant to a query, clusters these documents by subject, and returns summary of a cluster with the corresponding list of documents. This was originally developed for newswire documents, but has been used on Medline abstracts too. Our approach differs from QCS in that our focus is on terse cluster label generation that is indicative of the content of the cluster, rather than produce multi-document summary.

6 Conclusions and Future Work

We proposed a strategy for deriving sentence fragments from documents text that can serve as a cluster label for result set of a search query, specifically in the context of News documents involving queries centered around entities, events, and their relationships. The cluster label also provides additional information that serves as "answer" or "missing detail" relevant to the query. Even though our technique abstracts the meaning of a sentence as a set of stems of significant words in the sentence, it nicely incorporates preference for shorter labels and resuscitates word sequencing information for the label eventually. Thus, the final cluster labels are adequate, informative, and grounded in the document text, even though there are several examples in which they seem rather long. This approach also has potential to serve as a framework for generalizing or specializing clusters into an hierarchy. Currently, we are studying reliability, efficiency, and scalability of the entity-event timeline visualization application.

7 Acknowledgements

We wish to thank Don Loritz for enlightening discussions throughout this project.

References

1. O. Alonso and M. Gertz: "Clustering of Search Results using Temporal Attributes", *Proceedings of 29th ACM SIGIR Conference*, 597-598, 2006.
2. F. Beil F., M. Ester, and X. Xu: "Frequent term-based text clustering", In: *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining* (KDD-2002), 436-442, 2002.
3. R. Campos and G. Dias: "Automatic Hierarchical Clustering of Web Pages", *Proceedings of the ELECTRA Workshop with 28th ACM SIGIR Conference*, 83-85, 2005.
4. G. M. Del Corso, A. Gulli, and F. Romani: Ranking a stream of news, *Proceedings of 14th International World Wide Web Conference*, 97-106, Chiba, Japan, 2005.
5. D. Dunlavy, J. Conroy, and D. O'Leary: QCS: A Tool for Querying, Clsutering, and Summarizing Documents, *Proceedings of HLT-NAACL*, 11-12, 2003.
6. B. C. M. Fung, K. Wang, and M. Ester: "Hierarchical document clustering", in John Wang (ed.), *Encyclopedia of Data Warehousing and Mining*, Idea Group, 2005.
7. P. Ferragina and A. Gulli: "The anatomy of a hierarchical clustering engine for web-page, news and book snippets", In: *Proceedings of the 4th IEEE International Conference on Data Mining*, 395-398, 2004.
8. P. Ferragina and A. Gulli: "A personalized search engine based on web-snippet hierarchical clustering", *Proceedings of 14th International World Wide Web Conference*, 801-810, 2005.
9. A. Gulli: "The anatomy of a news search engine, *Proceedings of 14th International World Wide Web Conference*, 880-881, 2005.
10. B. J. Gao, and M. Ester: "Cluster description formats, problems and algorithms", *Proceedings of the 6th SIAM Conference on Data Mining*, 2006.
11. `http://www.google.com/trends`.
12. `http://www.lexisnexis.com/`.
13. `http://jakarta.apache.org/lucene/docs/index.html`.
14. `http://www.nlm.nih.gov/`.
15. S. Osinski, and D. Weiss: "A concept-driven algorithm for clustering search results", *IEEE Intelligent Systems*, May/June, 3 (vol. 20), pp. 48-54, 2005.
16. `http://vivisimo.com/docs/tagging.pdf`.