Emerging User Intentions: Matching User Queries with Topic Evolution in News Text Streams

Codrina Lauth
Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)
Sankt Augustin, Germany
Codrina.Lauth@iais.fraunhofer.de

Ernestina Menasalvas
Universidad Politecnica de Madrid
Madrid, Spain
emenasalvas@fi.upm.es

Abstract

Topic and event evolution analysis aiming at trend detection and tracking (TDT) from news data streams has considerably gained in interest during the last years. Consolidated studies have concentrated on identifying and visualizing dynamically evolving text patterns from news data streams. Detecting and understanding user behavior and relating user intentions to emerging topic trends in news data streams still continues to remain a huge challenge for making search engines in "real-time" responsive to user's information needs. In this paper, we will describe a three-layered approach *(user-system-content)* and how we can merge and process the relevant sources of highly evolving information on a news portal. This approach is a further step on building more streamline user-adaptive newswire search, based on topic and trend detection systems [9], able to deal in a user-friendly way with the permanently changing variety of news data streams that are embedded in the complex structure of a news portal.

Keywords: news streams analysis, query categorization, data mining

1 Introduction

Topic and trend detection from news data streams has grown considerably in interest on the commercial as well as on the research academic side over the last few years. According to recent surveys news browsing is one of the most important activities in the Internet. Several commercial products like the news engines, GoogleNews, YahooNews, MSNBot, Findory, NewsIn Essence have a huge number of users. Users are offered on several TV and radio broadcasting portals more and more complex newswire services, but the effective search in this highly dynamically evolving and content rich environments still remains a huge challenge. Looking at the structure of the news portals we see that three dimensions are permanently influencing one another: the user intentions are reflected in the news streams that appear on the site and these are embedded in the portal structure that is also changing according to the published content and user information needs. Consequently the challenge we want to approach here is to find a data analysis model that merges information from all three perspectives:

- the user intentions
- the system structure and
- the content of news streams.

Query categorization has attracted a lot of attention in research literature; see [3], [4], [18], [22]. The basic of our approach is grounded in the work [20] about enriched categorization of user queries related to meta information extracted from the portal structure. Here the authors have developed a two-stage approach of defining features of the user queries based on visibility and decay of the terms that are mapped in a taxonomy of concepts based on the site structure. The approach presented in this paper is extending the joint user-system scenario from [20] with one more dimension that we call the *content perspective*, where we try to find a model that matches the trends in user intentions on a news portal with the contents of the news streams.

The research in trend detection and tracking (TDT) systems has a long tradition. In the first comprehensive overview on TDT systems in

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 1725–1732
Torremolinos (Málaga), June 22–27, 2008

[17] an *emerging trend* in text streams is defined as a "*topic area growing in interest and utility over time*." The first TDT systems are exploring the *temporal evolution of usage patterns* [7] in text stream collections. But TDT applications dealing with emerging trends from the classical document-level perspective [2] have not been designed to respond to the immediate information needs of user queries on news portals. In our case, where we have to cope with highly dynamically information evolving environments, our concept of emerging trends from news data streams has to be re-defined from a *context-level*, being able to adapt intuitively to the *situated-user perspective* [9].

Up to our knowledge, there is no research work that combines the user and system perspective with topic and trend detection of content information from news data streams. Our major goal is create a processing scenario for automatically detecting and tracking emerging user information needs acting in a highly dynamically content and system environment. The analysis of user behavior on-the-fly, is helping us to detect their ultimate intentions and information need on the news portal. This will give us important hints on how the system can integrate the news in the portal structure intuitively so that the requested information is retrieved with higher accuracy. This overall *user-system-content approach* has motivated our present research in which we intend to automatically generate *emerging user query intentions* on a news website by enriching them with meta information hidden in the system of the website structure, and relating these results to the content information from the emerging news data streams.

Our paper has following structure: In the first section we present briefly our three fold *user-system-content* approach. The second section describes the three steps of our approach into more detail including the unifying context perspective. The third part contains a description of an application of our approach to a real news portal, including the mapping and implementation of the user and system perspective. The last part is dealing with the conclusion and future research.

2 User-System-Content Approach

News streams as highly evolving information units pose different challenges to text mining than usual text collections [11], because it has different characteristics, e.g. news can change unpredictably, it has an impact on other news and on the user behaviour on the portal, who are looking for these news, etc.[21] On the other hand, the mass of users is implicitly influencing which topics are becoming emerging trends on the news portals, because each portal as content provider is keen on adjusting its information services of the information needs of the target users they are trying to attract. Consequently, news has an important social-temporal dimension related to it, and we have to analyse news streams from the *context modeling perspective* [9], that is the unified user-system-content approach.

Figure 1 below is illustrating our three-layer approach. The system modeling perspective has to be merged to the user modelling perspective, as well as to the content modelling perspective, in order to obtain an overall *context model* able to deal with all relevant facets of a real-time, user-adaptive newswire services providing platform.

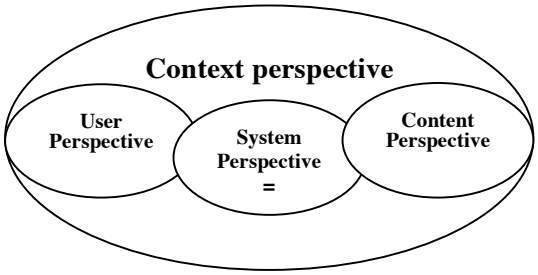


Figure 1: Defining emerging trends from the context perspective

From a ubiquitous user modeling perspective [9], our ultimate goal for designing an ongoing user-oriented news search functionality on a news portal is to automatically merge the content-based model of emerging trends from news data streams over time, to the user profiles, as well as to the web usage model derived from the prior user behavior extracted from the web query logs.

In the following sections we will describe each of the steps of our approach into more detail.

2.1. Analyzing user behavior *(Step 1: Integration of the user perspective)*

Staring from the user behavior analysis, reflected by the user queries logs, we want to match these results with the topic evolution in news data streams. Our first goal is to detect the user intentions from the user queries logs. The query terms are extracted from the web query logs and clustered according to the topics appearing in the headlines of the portal or of the latest news streams.

One of the main problems of categorization has to do with having a set of queries previously categorized. We investigate in the first step how to build an incremental learning model that allows us to classify queries in real time. Taking into account that words alone are not enough for a good categorization of query terms, we proceed on defining some visibility measures like in [20]. These measurements try to measure the degree in which the query submitted is similar to recent queries and/or contains terms that appear in the front page.

In [2] the first sets of measurements are called:

- *visibility front page_x*, where x deals with the interval of time in which the measure is taken;
- visibility to measure whether the query contains words of the front page during the last X days and/or whether it contains the most frequent terms in queries submitted in the lat X days …

Our approach is basically extending the previous two-fold method (user-system) proposed by [3] that is basically categorizing in a first step some queries according to the results they return and a fast classifier is built based on these results. In a second stage, which is described in our paper under section 2.2., it exploits properties of the visibility of the terms in the queries and in the front page along a period, to obtain a set of attributes that are used to further cluster queries and enrich classification.

2.2. How to build taxonomy of concepts reflecting the portal structure? *(Step 2: The mapping system perspective)*

Another important dimension that has to be integrated in our approach is the system modelling dimension, meaning the modelling of the portal structure in which the content information is embedded. We mean by system modelling dimension, a model that extracts and integrates semantic features from the metatags of each webpage as well as content features from the single news streams documents embedded in the webpage structure.

Looking closer at the structure and types of information on news portals, we find dynamically evolving news streams documents from multiple, multimedia sources, like podcasts/ video news, audio news, text streams news, link collections of headlines of different news. These different types of news streams are published on different sites on a news portal, and they can be sometimes content related. For a search engine in this environment it is hard to cope with the dynamicity of the content and structure change in an effective way.

Nevertheless, the portal itself can serve as an integration platform for all these heterogeneous types of information. The user query intentions (in 2.3.) have to be mapped to the evolving system environment integrated in the meta information of the website structure.

The system perspective of our approach is integrating the different types of news resources on the entire portal, for example from the “flow” type of news texts that come to the portal front sites constantly and regularly, at a rather fast pace, up to the “stock” type of text streams (e.g. headlines, or mainly the static web pages) that change unpredictably their formats and contents. [5].

We are investigating how clustering methods like PLSI and LDA (see a comparison of this two methods in [10]) can help to build the system model that will combine structure-related and document-related features.

2.3. Topic and trend detection from news *(Step 3: Integrating content perspective)*

We have analysed and detected the user intentions and related them to categories marked by the portal structure. Now the next step is to investigate how we can allocate the unstructured information hidden in the news streams to the user-system perspective in the most effective way. For this, we have looked at several approaches and tools dealing with trend detection and tracking (TDT) from text collections that are mentioned in the next section.

In the trend detection and tracking (TDT) research we find the distinction between "topic", "story" and "event" [17] of a news stream. In our approach we are not dealing at the moment with such fine-grained distinction of news streams document types. Here we are dealing mainly with *topics* that are higher-level, abstract terms used as labels to categorize collections of news streams, e.g. POLITICS, SPORTS, HISTORY, SCIENCE, etc. Topics can be derived into hierarchies of subtopics. Large news agencies are using in this context standardized taxonomies to categorize news topics, for example the standard IPTC NAA format for dpa news [23]. In general all these distinctions and different types of highly evolving semi-structured, as well as unstructured information make the modeling of retrieval process on a newswire portal very complex.

2.3.1. A selection of trend detection and tracking (TDT) systems

Several trend detection and tracking (TDT) commercial, as well as academic systems have been developed and studied over the last two decades. A detailed overview of these systems is available in [17, 5 and 8]. We will refer to some of these systems in this chapter, by illustrating this selection from the user perspective dimension that is relevant for our approach.

TDT systems can be automatic and semiautomatic [17]. The semiautomatic TDT systems need users input, in order to detecting trends of interest in a text stream collection. One of the first and most powerful TDT visualizing systems, that can be used also to visualize trends in news streams, is ThemeRiver (Havre et al. 2002 in [8]). Most of the classical document-based filtering TDT systems like ThemeRiver, TAO system, CIMEL, TimeMines, PatentMiner, HDDI, etc [8] are using "time-based keyword extraction techniques" [8] for trend detection from text streams. Depending on the types of text streams and the application environment, in which they are implemented, it is possible to extract other types of information. Commercial products like Autonomy/ClusterizedTM, SPSS/LexiQuest ClearForest/ClearResearch and ClearSight [8] show the same functionalities patterns like the academic systems mentioned above.

These systems have in common that they model well the time dimension of emerging trends from text stream collections, but they have not been designed to integrate the ongoing context dimension that is so time critical for user-adaptive newswire services providing platforms.

2.4. Step 4: The unifying context perspective

The initial goal of our three-fold approach is to define a user-system-content based keyword extraction and matching system based on visibility measures for topics in news streams, as enriched input for our trend detection and tracking on a news streams portal.

In this section we attempt to describe how we can integrated our previous three approached under one overall perspective, that we call the "context perspective" (see Figure 1).

Context and situation of the user are intimately connected. We define under the context perspective the combination of all three previous perspectives: user, system and content perspectives. This enriched view is capturing all relevant information for defining the features relevant for evolving user intentions.

According to [16] a situation consists of two parts: the user-related factors that are

intrinsically tied to a user as abilities, goals or personal traits and the environment in which the user perceives and acts that can be described by factors that are independent of an individual user. In our case the user-related factors are approached by the user perspective that will be described section 3.1. and the environment is marked by the system perspective, which will be elaborated in the next section.

In the following section, we will show a use case on how we analyse user behaviour and how we can match the user query results to the structure of a news portal and to different content levels of a news stream.

3 The application of our approach to a real news portal

We present a possible implementation of the presented approach on a real news site that uses GSA as search engine. For the exemplary case analysis of a real news website, we could not include the concrete mapping of the content perspective at this stage, since we are just evaluating the TDT systems matching best our approach.

The initial goal of our three-fold approach is to define a user-system-content based keyword extraction and matching system based on visibility measures for topics in news streams, as enriched input for our trend detection and tracking on a news streams portal.

From the previous studies [20] we will related to the following results: weblog used corresponds to the activity of the site along 4 months containing a total of 442909 records. After cleaning and pre-processing the log,106443 different queries have been identified in which a total of 43742 different terms are used.

3.1. The user perspective

Consequently we define the visibility of a term in a period of time V (t; p) based on its appearance frequency (number of times it appears) in the period. According to this coefficient we define:

- stable terms in a period st(p): those terms t that appear along the period p under study in uniform way. Uniformity is a function of the average number of appearance of the terms in the period. It also depends on the decay function of the term d(t; p) that reflects the ratio of appearance along time. Values of decay over 1 means increasing ratio of appearance, values close to 0 means decreasing ratios and values around 1 means stable term
- new term nt(p): a term t belongs to this set if it appears for the first time in the period p
- top-ten: one term t belongs this set if in the period p appears over the average ratio of appearance

All these sets are calculated for different period length: 24 hours, 7 days, 31 days and both for the appearance of words in heading in the front page and in queries. Queries and later sessions are enriched with information regarding the terms they contain: 24 hours top ten, preceding day, week before. This attributes will be used for the enriched categorization to deal with dynamism of words and users.

3.2. The system perspective

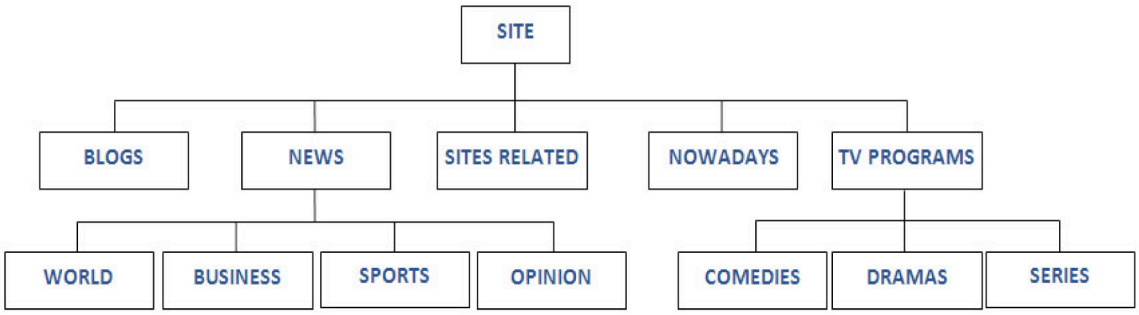
We propose to deal with two kinds of categories reflecting the structure of the site:

- The one inherent in the site structure: this is very stable taxonomy as the categories reflecting the structure of the site does not change very often
- The one extracted from the front page: this is a very changing taxonomy, because it depicts only the categories that are currently of interest for the largest mass of users, thus it reveals interesting information regarding the dynamics of the users and their interests

We use these two taxonomies to cluster the news streams documents, according to the their appearance on the different sites. This means that each news stream will be marked with

labels from the two taxonomies. In Figure 2 we have an example of a taxonomy extracted from the front page of a TVBroadcast news website.

Figure 2: Headlines defining the site structure



3.3. Mapping

Interestingness of the query highly depends on features of the terms contained in the query. Consequently we firstly categorize terms and then according to the label queries. Two criteria are taken into account:

- visibility of the term on the queries
- visibility of the term on the front page

3.4. Implementation

The work specified in this section is inspired by the results of [20]. According to the assumptions mentioned in [20] on how fast-feature classifier works, the method we propose is similarly composed of the following stages:

1. Off-line pre-processing:
   - Taxonomies creator
   - Web Log cleaning: this process has to be repeated due to the stream nature of the log data
   - Visibility sets calculation: as in the previous step, sets has to be updated as new data arrived and are cleaned
2. Manual labelling of queries
   - Search sets of results for a selected set of queries: Randomly some queries are chosen and are sent to the search engine to analyze the set of results.

Mapping of the results into domain taxonomy. We propose to use only the urls of the result set for the mapping in the taxonomy as in [3].

3. Front-page labelling of queries. For those queries containing terms of the headings in front page, randomly some queries are chosen and manually labelled
4. Fast-feature classifier: We call it fast-feature as only the words contained in the query are used to build the classifier
   - Reduce calculation. Sets of words that appear in the categories chosen are used so to decide which of them are distinctive for each category. We use rough set based techniques for this step.
   - Rough Sets based classifier. Based on the results we propose to build the classifier
   - Build final model
5. Enrichment of the queries regarding to visibility functions
6. Enriched classifier
7. Evaluation

4 Conclusion and future research

We have described into detail the three layered approach *user-system-content*, which reflects the "matching process" from both sides of the coin, namely how topic evolution in news streams are affecting the user behavior over time (topic-to-user view) and how can we optimize the matching of user query results to an emerging stream of news texts (user-to-topic view).

Matching emerging user intentions (i.e. trends in user behavior) with highly dynamically evolving

information (like e.g. trends in news streams) helps us to define user-adaptive visibility measures for emerging trends in news streams.

The proposed approach is only the first starting point on how to optimize the matching of user query terms to the emerging topics extracted from the stream of news texts by using the structure of the portal where they are located.

Our next steps will lead us to find more effective ways of modelling and optimizing the synchronization process of trends detection and user behaviour influencing the topic evolution in news streams, over time by taking into account the site location of news streams.

Acknowledgements

For this paper we acknowledge the support of the EU-funded project KDubiq-CA (IST-6FP-021321, www.kdubiq.org) under FET Open (Future and Emerging Technologies) Unit.

References

1. J. Allan (2002). Topic Detection and Tracking. Kluwer Academic Publishers.
2. S. M. Beizel, E. C. Jensen, A. Chowdhury, D. Grosman, O. Frieder (2004). Evaluation of Filtering Current News Search Results. *In Proceedings of SIGIR 2004.*
3. S. M. Beizel, E. C. Jensen, A. Chowdhury, D. Grosman, O. Frieder (2007). Automatic classification of web queries using very large unlabeled query logs. *In Proceedings for ACM Trans. Inf. Syst., 25(2):9.*
4. D. Bollegala, Y. Matsuo, M. Ishizuka (2007). Measuring semantic similarity between words using web search engines. *In WWW '07: Proceedings of the 16th international conference on World Wide Web*, ACM Press, NY, USA, p. 757/766.
5. K. K. Bun (2004). *Topic Trend Detection and Mining in World Wide Web*. Dissertation at Tokyo University, Khoo Khyou Bun, 2004.
6. G. M. del Corso, A. Gulli, F. Romani (2005). Ranking a Stream of News. *In Proceedings of the www Conference in 2005.*
7. A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Schneiderman, C. Plaisant (2007) Discovering interesting usage patterns in text collections: integrating text mining with visualization. I*n Proceedings of the sixteenth ACM conference on Conference on in*formation and knowledge management, 2007, p. 213 - 222
8. M. Dubinko, R. Kumar, J. Magnani, J. Novak. P. Raghavan. A. Tomkins (2006). Visualizing Tags over Time, *In Proceedings of WWW2006.*
9. D. Heckmann (2005). *Ubiquitous User Modeling*, Dissertation at the University of Saarbrücken, Germany.
10. G. Heinrich, J. Kindemann, C. Lauth, G. Paass, J. Sanchez-Monzon (2005). Investigating Wordcorrelations at Different Scopes – A latent-Concept Approach. *In Workshops Proceedings of the ICML05, Bonn, Germany.*
11. S. Henning, M. Wurst (2006). Incremental Clustering of Newsgroup Articles. IEA/AIE 2006: 332-341
12. M. Henzinger, B.-W. Chang, B. Milch, S. Brin (2003). Query-Free News Search.WWW2003, Budapest.
13. A. Kaban, M. Girolami (2002). A dynamic probabilistic model to visualize topic evolution in text streams. J. Intell. Inf. Syst. 18(2-3): 107-125.
14. S.-K. Kim, K.-C. Lee (2007). Trend Analysis Using a Temporal Web Ontology Language in News Domains. *In COMPSAC 2007 IEEE.*
15. J. Kleinberg (2006). Temporal Dynamics of On-Line Information Streams. In M. Garofalakis, J. Gehrke, R. Rastogi (Eds), *Data Stream Management: Processing High-Speed Data Streams*, Springer, 2006.
16. Kray (2003). *Situated interaction on spatial topics*, In Volume 274, Dissertations in Artificial Intelligence, In_x November.
17. A. Kontostathis, L. M. Galitsky, W.M. Pottenger, S. Roy, D. J. Phelps (2004). A Survey of Emerging Trend Detection in Textual Data Mining. *In Survey of Text Mining-Clustering, Classification and Retrieval* - by Michael W. Berry, 2004.

18. B. Kules, J. Kustanowitz, B. Shneiderman (2006). Categorizing web search results into meaningful and stable categories using fast-feature techniques. *In JCDL 2006*, p.210-219, 2006.

19. M. Maslov, A. Golovko, I. Segalovich, P. Braslavski. Extracting News-Related Queries from Web Query Log. from the website under URL: *http:// company. yandex.ru/artcles/news-related-queries.xml*

20. E. Menasalvas, S. Eibe, M. Valencia, , P. Sousa (2007). An integrated Web-Query Classification based on Rough Sets. *In AWIG 2007 Proceedings.*

21. M. Montez-y-Gomez, A. Gelbukh, A. Lopez (2002). Mining the News: Trends, Associations and Deviations.

22. D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, Q. Yang (2006). Query enrichment for web-query classification. *In Proceedings of ACM Trans. Inf. Syst.*, 24(3):320-352.

23. http://www.iptc.org/pages/index.php