The Anatomy of a News Search Engine

A. Gulli
Dipartimento di Informatica, University of Pisa
gulli@di.unipi.it

ABSTRACT

Today, news browsing and searching is one of the most important Internet activity. This paper introduces a general framework to build a News search engine by describing Velthune, an academic News search engine available on line.

Categories and Subject Descriptors

H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Design, Experiments

Keywords

News Search Engines, Information Extraction, Syndication

1. INTRODUCTION

According to a recent survey [1] made by Nielsen/ NetRatings, news browsing and searching is one of the most important internet activity. In October 2004, there were more than 28 millions of active news users in U.S.. For instance, Yahoo! News had an audience which is roughly the half of Yahoo! Web Search, a third of Google Web Search and a bit more than AOL Web Search. The huge amount of news available on line reflects the users' need for a plurality of information and opinions. News Search engines are then a direct link to fresh and unfiltered stream of information. There are many commercial News search engines. Google News retrieves news information by more than 4,000 sources, organizes it in categories and automatically builds a Web page with the most important news for each category. Yahoo news runs an analogous service on more than 5,000 sources. Microsoft recently announced its NewsBot, a news engine that provides personalized news according to different profiles built for each users. Findory proposes a similar personalized service, which relies on patent pending algorithms. A list of commercial news engine is given in [2].

Despite this great variety of commercial solutions, we found just few academic papers on this subject. NewsInEssence [7] is a system for finding and summarizing clusters of related news articles. QCS [8] is a software tool for streamlined IR from generic document sets. [5] proposes a topic mining framework for news data stream. [11] finds news articles on the web that are relevant to TV news currently being broadcast. [12] proposes a tool to automatically extracting news from Web sites. NewsJunkie [10] is a system that personalizes news for users by identifying the novelty of stories in the context of stories users have already reviewed. We think that the few scientific publications cause the News search engine technology to remain largely a black art.

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

2. A NEWS SEARCH ENGINE: VELTHUNE

In this short paper, we introduce a more general framework to build a News search engine. Our system, Velthune, is a complete News search engine for retrieving, ranking, indexing, classifying, clustering and delivering personalized news information extracted both form the Web and from news feeds. The system is made by the modules depicted in Figure 2. It has a Web interface at http://newsengine.di.unipi.it/. In the following, we describe them.

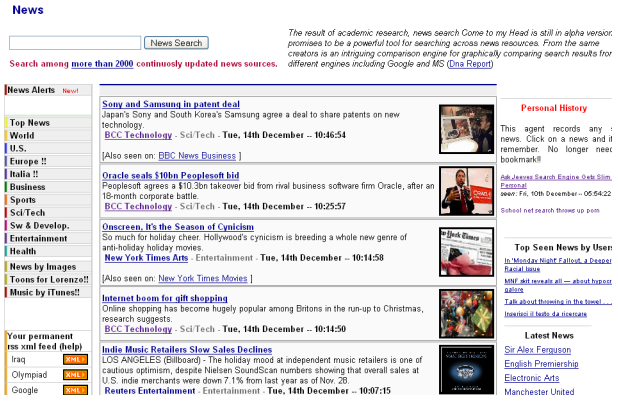

Figure 1: Velthune Web interface.

The retrievers: This module gathers news from a set of selected remote news sources. It supports different syndication formats, such as Atom [3] and RSS [4]. Besides, it is possible to feed the system with selected information extracted from remote Web pages. Currently, we selected a list of news sources consisting of about 2000 different Web journals. For efficiency reason, the space of news sources is partitioned and this module is composed by several processes which run in parallel. The data is collected 24h per day, and the stream of information is stored into a local database $\mathcal{N}_{\mathtt{db}}$.

The images processor: The image processor module analyzes information stored in the $\mathcal{N}_{\mathtt{db}}$. It tries to enrich any news with an associated image. In the easy case (e.g. Toons category), the news source has already associated an image to a given news in the RSS/Atom feed. This association is expressed as an HTML fragment, so we can easily download

the image locally and create a suitable thumbnail of it. In other situations, we have a news n, extracted by a Web page p or by a RSS/Atom feed,with no associated image which refers to a Web page p. We download locally any image contained in p and use many heuristics to identify the most suitable image to be associated to n. We take in account contextual information (e.g. where the image is placed in the HTML source) as well as content information (e.g. image's size, colors and other features).

The feature selection index: We use a the feature selection index $\mathcal{F}_{\text{db}}$ to design an efficient method for extracting meaningful terms from the news. As suggested in [9], we index the whole DMOZ and implement a `TFxIDF` measure that is *DMOZ-category* centered. $\mathcal{F}_{\text{db}}$ is built at preprocessing time and then used for selecting and ranking on-the-fly key features used for News classification and clustering.

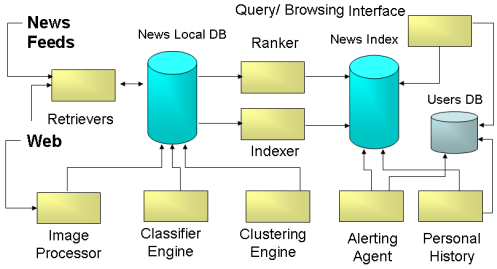

Figure 2: Architecture of the News Engine.

The classifier engine: All the news in $\mathcal{N}_{\text{db}}$ needs to be classified. The categories defined in the system are given in figure 3. We use a naive bayesan classifier. Note that a (relatively large) part of the RSS/Atom feed are already manually classified from the originating news source. As a consequence, the key idea for classifying is to use the classifier in a *mixed mode*: as soon as an already classified news by a news source is seen, the classifier is switched in training mode; the remaining unclassified news are categorized with the classifier in categorizing mode.

Category	# News	Category	# News
Business	57079	Entertainment	71865
Europe	29256	Health	20025
Italia	13098	Music Feeds	1112
Sci/Tech	40368	Software & Dev.	2739
U.S.	13368	Top News	88189
Sports	62672	World	80323

Figure 3: News collected over three months.

The clustering engine: We adopt a variant k-means algorithm with distance threshold for creating a *flat clustering* structure, where the number of clusters is not known a-priori. This is much useful for discovering similar news or news mirrored by many sources.

The ranker engine: Ranking news is a task rather different than Web page ranking. From one side, we can expect a less amount of spam since the news come from controlled sources. From other hand, when a news is posted it is a fresh kind of information. Therefore, there is almost no hyperlink pointing to it. In a companion paper [6], we describe a News ranking algorithm which ranks both news and sources, and takes in account many different factors such as: (1) news freshness (2) news clustering aggregation (3) importance of the source posting the news. Our running prototype have used a ranking algorithms based on (1) and (2) for months. Currently, we are integrating other criteria suggested in [6].

The indexer: Classified and Clustered news coming from $\mathcal{N}_{\text{db}}$ are indexed by this module, which produces an inverted list index. For each news we store the source, the url, the title, the associated category, the description, a rank value, and the publication date.

Query and Browsing interface: The index is accessed by the search and browsing interface. A typical user can access the news by category, search within any category and search all the indexed news. All the results are available with a public Web interface. Besides, we provide an RSS search mode. Using this feature, a user can submit a list of keywords and she receives a permanent XML feed with the most updated news about that topic. In this way, Velthune acts as a public aggregator for news feeds.

Personalized alerting agent: We integrated a mail alerting system, which produces a personalized daily summary of the news chosen by users. In fact, each user can login into the system and record a private set of queries. As soon as a fresh and related news appears into the system, it is aggregated into a mail sent daily to the user. Information about users is stored in a local database $\mathcal{U}_{\text{db}}$.

Personal history tracker: This module allows a transparent tracking of the the news selected by the users through different browsing or searching sessions. Currently, we use this feature for alleviating users from the need to bookmark interesting news. Data are stored in $\mathcal{U}_{\text{db}}$ and are kept private. Therefore, they are accessible preserving the users' privacy. In this case, the system does not request an explicit login, but works using a cookie mechanism. We also plan to use this kind of data in an anonymized form, for providing personalized search results.

Helper modules: Velthune provides other services available trough the Web interface. They are: (i) "*Top Ten*", the most accesses news for each category; (ii) "*Top News Sources*", the most active news sources for each category; (iii) "*Latest News*", the list of most frequent named entities dynamically extracted at query time from each category.

3. REFERENCES

[1] `http://www.nielsen-netratings.com/`.
[2] `http://searchenginewatch.com/`.
[3] `http://www.atomenabled.org/`.
[4] `http://blogs.law.harvard.edu/tech/rss`.
[5] S. Chung and D. McLeod. Dynamic topic mining from news stream data. In *ODBASE*, 2003.
[6] G.M. Del Corso, A. Gulli, and F. Romani. Ranking a stream of news. In *www14*, 2005.
[7] S. Blair-Goldensohn D. R. Radev, Z. Zhang, and R. S. Raghavan. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *HLTC*, 2001.
[8] D. M. Dunlavy, J. P. Conroy, and D. P. O'Leary. Qcs: A tool for querying, clustering, summarizing documents. In *HLT*, 2003.
[9] P. Ferragina and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering. In *www14*, 2005.
[10] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *www13*, 2004.
[11] M. Henzinger, B. Chang, B. Milch, and S. Brin. Query-free news search. In *www12*, 2003.
[12] D. Reis, P. Golgher, A. Silva, and A. Laender. Automatic web news extraction using tree edit distance. In *www13*, 2004.