Building a Digital Library of Web News

Nuno Maria, Mário J. Silva

Faculdade de Ciências da Universidade de Lisboa, Portugal
Departamento de Informática
{nmsm, mjs}@di.fc.ul.pt

Abstract. We introduce a new information system for organization of a Digital Library of news articles found on the Web, with automatic topic classification. We present our strategies to deal with different update frequencies of news Web sites, the classification methodology, the data model for storing news articles, measurements on the data retrieved and finally results of classification of this type of information.

Overview

The number of publications and news articles published on the Web had a dramatic increase over the last years. Readers often want to search and retrieve these publications for past news articles related to a particular subject or story. However, a Web news archive is usually not available and we can not rely in the publication archive. Search engines fail in coverage and often present broken links to publications. In addition, electronic publications separate their articles into a set of categories, but classifications are not uniform, leading to a poor satisfaction of readers information needs. Our project involves the creation of a framework to define Web services that let users see past published news on the Internet sites organized in a common category scheme.

To achieve this, specialized information retrieval and text classification tools are necessary. The Web news corpus suffers from specific constraints, such as a fast update frequency or a transitory nature, as news information is “ephemeral.” As a result, traditional IR systems are not optimized to deal with such constraints. As each publication has its own scheme of topics, it is also difficult to match the classification topics defined by each publication.

Our framework for Web news retrieval is built on broadly available research work [1]. In the automatic text categorization field, detailed examinations on the behavior of statistical learning methods and performance comparisons are periodically available [2, 7]. These studies show Support Vector Machines as the most efficient technique currently available. Recent work on Topic Detection and Tracking (TDT) and document clustering is also available [8]. These fields are studying automatic techniques for detecting novel events from streams of news stories and track events of interest over time. However, these areas are still a recent research activity and many research questions remain open.

In our work, we are addressing some of these problems applying advanced information retrieval and classification techniques to the physical world of Web publishing, as we index and classify the major Portuguese news publications available on the Web.

In our research, we have identified the following main problems associated with retrieval of theme-based news:

- In general, news articles are available on the publisher's site only for a short period of time. Many Publications do not give access to their archive of previous editions and a database of references becomes easily invalid.
- Many news Web sites are built dynamically, often showing different information content over time in the same URL. This invalidates any strategy for incremental gathering of news from these Web sites based on their address.
- Direct application of common statistical learning methods to automatic text classification raises the problem of non-exclusive classification of news articles. Each article may be classified correctly into several categories, reflecting its heterogeneous nature. However, traditional classifiers are trained with a set of positive and negative examples and typically produce a binary value ignoring the underlying relations between the article and multiple categories;
- News clustering, which would provide easy access to articles from different publications about the same story, can be an important improvement. The automatic grouping of articles into the same topic requires very high confidence, as mistakes would be too obvious to readers.

To address the above presented problems we believe that it is necessary to integrate in a global architecture a specialized retrieval mechanism and a multiple category classification framework, including a data model for information and classification confidence thresholds.

We have been developing and evaluating a prototype of a system that addresses the above introduced requirements.

Architecture

Our system integrates a set of built or customized components for retrieval and classification of text documents. Figure 1 presents our system architecture for retrieval and classification of Web news.

The Retrieval Framework is implemented as a modified version of the Harvest System [1]. To deal with multiple publication periodicities, several queues are created and managed, expressing different update priorities and different gathering schedules. Information mediators [6] specialized for each publication are configured with rules to determine the URLs of the current edition and perform the collection of news articles according to the different schedules of each scanned publication. On a higher abstraction level, we view the Retrieval Framework as a service provider of a continuous stream of news articles retrieved from news Web sites. Each retrieved article is then delivered to our Classification Framework.

As articles arrive from the stream provided by the Retrieval Framework, they are automatically converted into a vector format according to a pre-defined vocabulary.

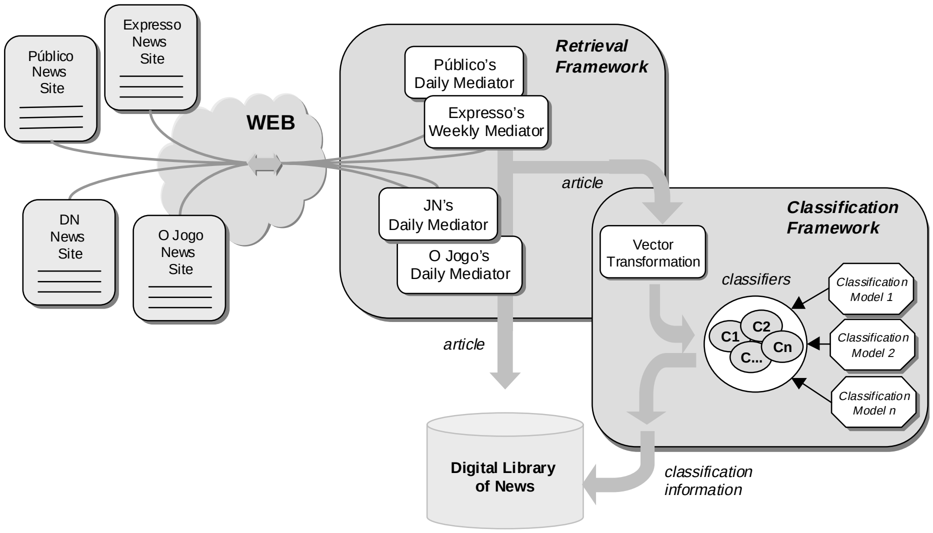

Fig. 1. Architecture for retrieval and classification of news articles. The Retrieval Framework collects articles from news sites distributed over the Web by specialized information mediators. These articles are sent to the Classification Framework for classification on a common scheme according to train examples provided by a local Digital Library. The generated classification information is then stored in our Digital Library of News for use by publications

The actual classification, based on pre-computed classification models, one for each pre-defined category, is performed with this vector format. The models are built with trained examples prepared from the news corpus provided by the local news library [4]. In our implementation this corpus is actually a daily newspaper with manually classified articles. The Classification Framework uses SVM^{ligth}, a package developed for automatic text classification using Support Vector Machines [3].

Once one article is processed, its vector format and classification are loaded in our Digital Library of News. It also stores the article's confidence level, returned from the classifier for each model, and possibly related articles. Proximity between articles is detected by a Similarity Detection mechanism. Each new article is compared to recently processed articles to check for similarities. If similarity is detected, then the article is grouped with other articles on a related cluster.

Results

The Portuguese online news corpus has 15 national news wires. We validated our system with a sample of 5 publications. Its overall performance on this sample is summarized in Table 1. We estimate in 4 GB annual storage requirements to maintain the text of the 5 publications. We believe these are acceptable storage needs considering today's hardware cost. Our strategies, restricting mediators to index only online "interesting" content, provide scalability to our solution making it a useful tool to small communities with specific interests.

Table 1. Measurements made in our Digital Library of News with 5 publications

Average Daily insertions	Average text+metadata size	Average text size	Average classification time
762 articles	12.0 KB / article	4.4 KB / article	1.0 sec / article

Our Classification Framework was validated with a set of articles from the publications sample representing a day of publishing activity (approximately one thousand articles). Here, we only outline the main results. A more detailed discussion is available in [5].

The classification mechanisms achieved 94,5% of accuracy with non-exclusive classification. Approximately 37% of the articles were classified in more than one of the eleven defined categories.

In the extremely dynamic environment of Web news we also must be aware of the degradation of the classifiers' accuracy in time. We observed that classifiers should be trained with a set of articles from a wide temporal range to minimize the effects of season events. In our work, we also developed a framework for detecting deficient classifiers and minimize accuracy degradation.

We are now applying and measuring the accuracy of clustering techniques to our system.

References

1. Bowman, C., Danzig, P., Hardy, D., Manber, U. and Schwartz, M.: The Harvest Information Discovery and Access System. *Proceedings of the Second International WWW Conference*. pp.763-771, 1994.
2. Dumais, S., Platt, J., Heckerman, D. and Sahami, M.: Inductive Learning Algorithms and Representations for Text Categorization. *Proceedings of the Seventh International Conference on Information and Knowledge Management*, 1998.
3. Joachims, T.: Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
4. Maria, N., Gaspar, P., Grilo, N., Ferreira, A. and Silva M. J.: ARIADNE - Digital Library Architecture. *Proceedings of the 2nd European Conference on digital Libraries (ECDL'98)*, pages 667-668, 1998.
5. Maria, N. and Silva, M. J.: Theme-based Retrieval of Web News. *Proceedings of the Third International Workshop on the Web and Databases (WebDB'2000)*. To be published as Springer LNCS.
6. Wiederhold G.: Mediators in the Architecture of Future Information Systems. *IEEE Computer*, pages 38-49, March 1992.
7. Yang, Y. and Liu X.. A re-examination of text categorization methods. *Proceedings of the 22th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99),* pages 42-49, 1999.
8. Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald B. T. and Liu X.. Learning approaches for Detecting and Tracking News Events. *IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval*, 14(4), July/August 1999.