Theme-based Retrieval of Web News

Nuno Maria, Mário J. Silva
DI/FCUL
Faculdade de Ciências
Universidade de Lisboa
Campo Grande, Lisboa
Portugal
{nmsm, mjs}@di.fc.ul.pt

ABSTRACT
We present our framework for classification of Web news, based on support vector machines, and some of the initial measurements of its accuracy.

OVERVIEW

The number of publications and news articles published on the Web had a dramatic increase over the last years. Readers often want to access news articles related to a particular subject such as sports or business. Electronic publications already 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 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 watch the theme with the classification topics defined by each publication.

Our framework for Web news retrieval is built on broadly available research work. In the automatic text categorization field, detailed examinations on the behavior of statistical learning methods and performance comparisons are periodically available [5]. From that research, support vector machines appear to be the most efficient technique available. Recent work on the topic detection and tracking (TDT) and document clustering is also available [1, 6]. 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.

Text categorization applied to news information is a complex task, given the information's subjective and heterogeneous nature. We identified the following main problems:

- 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;
- It is necessary to validate the classification models created with the trained examples from our local digital library of news information. As each publication has a different classification scheme, manual labeling of validating test examples is sometimes necessary;
- Accuracy. The classification system must measure classification confidence and prevent misclassifications, as they have a strong negative impact in the reader. In Web news, good values for classification accuracy obtained in research environments (over 90%) may be insufficient. A reader that detects semantic incoherence will probably distrust the system and stop using it;
- News clustering, which would provide easy access to articles from different publications about the same story, is another risky operation. Grouping of articles into the same topic requires very high confidence as mistakes would be too obvious to readers.

To address the set of problems presented above, we provide a multiple category classification framework. We obtained measures and thresholds for classification confidence, to optimally integrate retrieval and classification components in a

global architecture. In our research, we apply new techniques to news clustering, which have not been tested and could produce higher levels of accuracy. We also want to study how classifiers behave as the time distance between the date of articles in the training set and evaluated articles increases. We intend to measure how they lose accuracy as new vocabulary is added and strong weight features are discarded.

RETRIEVAL AND CLASSIFICATION SYSTEM 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 Service is implemented as a modified version of the Harvest System [2]. To deal with different publication periodicity, several queues are created and managed expressing different update priorities and different gathering schedules. Gathering agents 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 service as a provider of a continuous stream of news articles retrieved from news Web sites. Each retrieved article is then delivered to Classification Service component.

As articles arrive from the stream provided by the Retrieval Service, they are automatically converted into a vector format according to a pre-defined vocabulary. 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 Service uses SVM^{ligth}, a package developed for automatic text classification using support vector machines [3].

Once one article is processed by the Classifier, its vector format and classification are loaded in the Classified Articles Database. 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 the

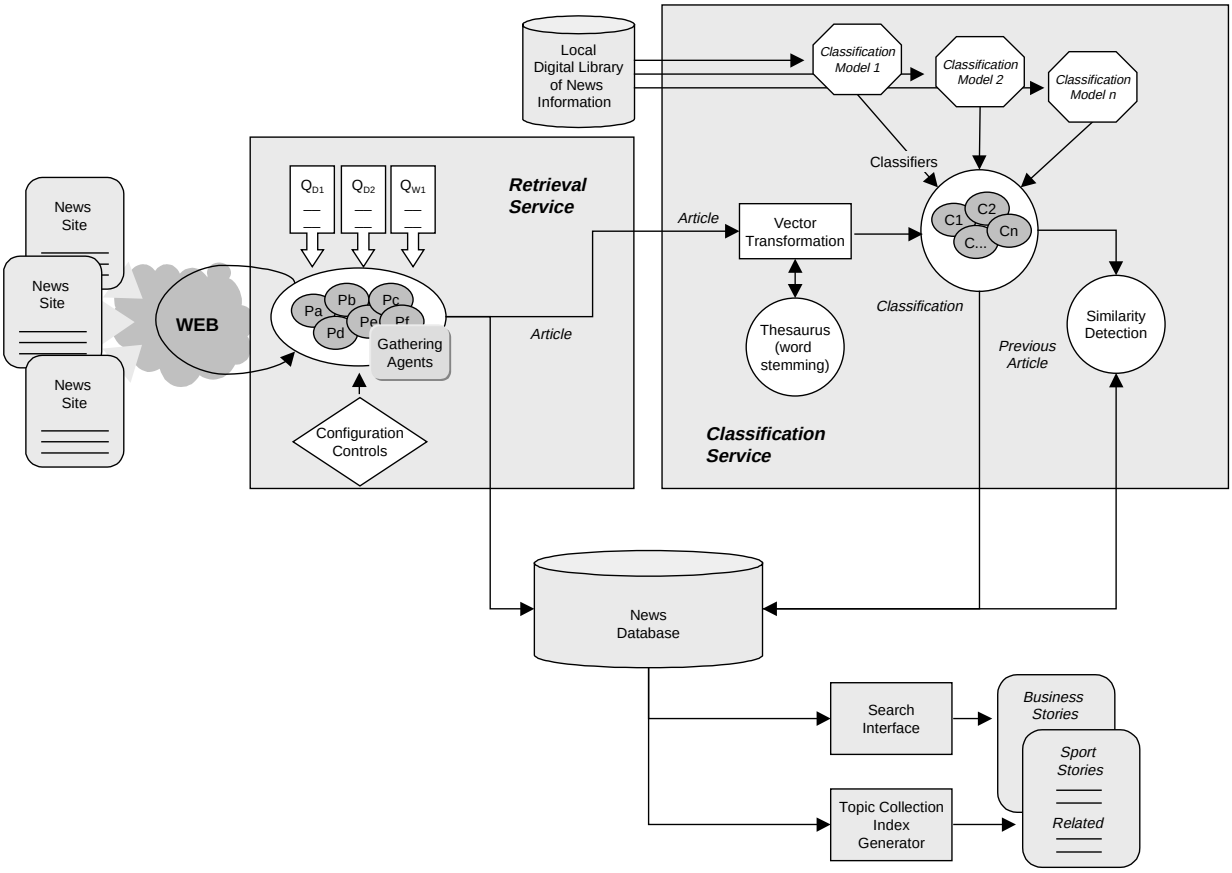

Figure 2. Architecture for retrieval and classification of news articles. The Retrieval Service collects articles from news sites distributed over the Web. These articles are sent to the Classification Service for classification on a common scheme according to train examples provided by a Local Digital Library. The generated classified information is then stored for use by publications. Topic Collection Index generators update topical Web portals.

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 news corpus has 15 national wide news wires with different periodicity. We validated our system with a sample of articles of 5 publications representing a day of publishing activity (approximately one thousand articles). Our classification mechanisms achieved 94,5% of accuracy with non-exclusive classification. Approximately 37% of the articles were classified in more than one category.

In the extremely dynamic environment of Web news we also must be aware of the degradation of the classifiers' accuracy in time. Figures 3a and 3b present the behavior of four category classifiers with two distinct training strategies. In the first experiment, we built models with articles from January and classified news dated from the following months. In the second strategy, we built models with selected articles from all the months of the year.

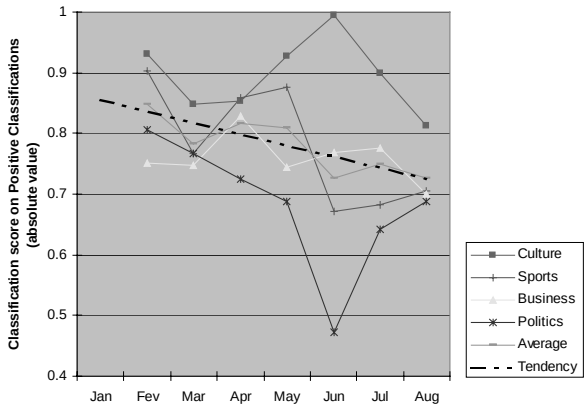

Figure 3a. The classification score on positive examples in time, obtained with models trained with articles from January. In topics with low seasonality such as culture, confidence is higher, but in other topics, like politics, confidence may decrease sharply.

The degradation of accuracy, measured by the confidence on positive classification, obtained by linear regression of the average behavior is evident with the first strategy. However, the reduced number of categories and our classifier building strategies do not force any feature selection, minimizing this effect.

Due to this degradation, our final classification models were built with selected articles, uniformly distributed along the year. This approach tends to minimize the negative effects of cyclic events.

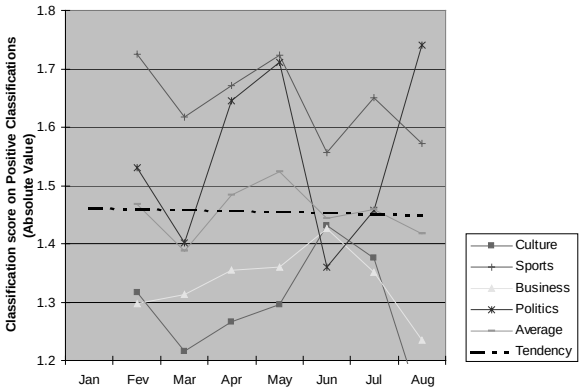

Figure 3b. The classification score on positive examples in time, obtained with models trained with articles of a full year. Confidence in classifications is significantly higher than with the previous strategy. Although some variations are visible, the overall behavior is more stable in time.

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

REFERENCES

[1] Allan, J., Papka, R. and Lavrenko V.. On-line New Event Detection and Tracking. In *Proceedings of the 21th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 37-45, 1998.

[2] Bowman, C., Danzig, P., Hardy, D., Manber, U. and Schwartz, M.. The Harvest Information Discovery and Access System. In *Proceedings of the Second International WWW Conference*. pp.763-771, 1994.

[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. In *Proceedings of the 2nd European Conference on digital Libraries (ECDL'98)*, pages 667-668, 1998.

[5] Yang, Y. and Liu X.. A re-examination of text categorization methods. In *Proceedings of the 22th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 42-49, 1999.

[6] 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*, Vol. 14(4), pages 32-43, July/August 1999.