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 introduce an information system for organization and retrieval of news articles from Web publications, incorporating a classification framework based on Support Vector Machines. We present the data model for storage and management of news data and the system architecture for news retrieval, classification and generation of topical collections. We also discuss the classification results obtained with a collection of news articles gathered from a set of online newspapers.

1. INTRODUCTION

The number of publications and news articles published on the Web had a dramatic increase over the last years. Existing keyword query-driven search engines cannot cope with this ever-growing mass of information. Readers often want to access news articles related to a particular subject such as sports or business. Many electronic publications on the Web already separate their articles into a set of categories, but classification criteria are not uniform, leading to a poor satisfaction of reader's 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." In addition information availability is also uncertain. As a result, traditional IR systems are not optimized to deal with such constraints.

Research work of relevance to our goal is broadly available. In the automatic text categorization field, detailed examinations on the behavior of statistical learning methods and performance comparisons are periodically available [9]. Bayesian probabilistic approaches, nearest neighbor, decision trees, inductive rule learning, neural networks and support vector machines are some techniques applied to this domain. However, and according to several recent experiments [3, 4, 9], support vector machines appear as the most efficient technique available. Its efficiency and accuracy are only compared with nearest neighbor learning approaches, and between these two and the others there is a big gap.

Recent work on topic detection and tracking (TDT) and document clustering is also available [1, 10]. These fields are studying automatic techniques for detecting novel events from streams of news stories, and track events of interest over time. However, these topics are recent research and many 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 gather, index and classify the major Portuguese news publications available online.

The Web news environment is very unstable: news information quickly becomes obsolete over time and is discarded by publishers. In addition, text categorization applied to news information is also a complex task, given the information's subjective and heterogeneous nature.

In our research, we have identified the following main problems associated with automatic 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 relationships 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 of the classification activity: 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. The automatic grouping of articles into the same topic requires very high confidence, as mistakes would be too obvious.

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

We present the design of a prototype of a system that addresses the above introduced requirements. We have used it to study how classifiers behave as the time distance between the date of news articles in the training set and evaluated news increases. We measured how they lose accuracy as new topics and vocabulary are added and strong weight features are discarded.

2. SUPPORT VECTOR MACHINES

Support Vector Machines - SVMs are a new learning method introduced by Vapnik [8], but only recently they have been gaining popularity in the learning community. In its simplest linear form, an SVM is a hyperplane that separates a set of positive examples from a set of negative examples with a maximum margin. Figure 1 illustrates this scheme.

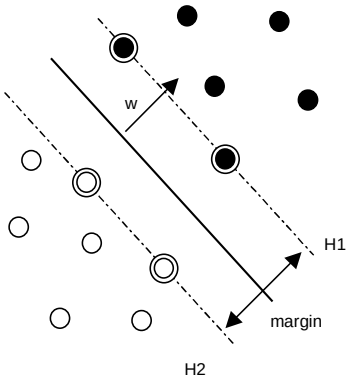

Figure 1. Linear separating hyperplanes. The support vectors are circled.

The formula for the output of a linear SVM is $u = \vec{w} \cdot \vec{x} - b$, where $\vec{w}$ is the normal vector to the hyperplane, and $\vec{x}$ is the input vector. In the linear case, the margin is defined by the distance of the hyperplane to the nearest of the positive and negative examples. Maximizing the margin can be expressed as an optimization problem: minimize $\frac{1}{2} \| \vec{w} \|^2$ subject to

$$y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, \forall i$$

where x_i is the ith training example and y_i is the correct output of the SVM for the ith training example.

SVMs have been shown to yield good generalization performance in a variety of classification problems, including handwritten character recognition, face recognition and more recently text categorization [4]. The simplest linear version of the SVM provides good classification accuracy, is fast to learn and fast classifying new instances.

3. THEME-BASED NEWS RETRIEVAL SYSTEM

Our system integrates a set of built or customized software components for retrieval and classification of text documents and a shared database of news articles. Figure 2 presents our architecture for retrieval and classification of Web news. It has the following main components:

Retrieval Service: This Service was implemented as a modified version of the Harvest Information Discovery and Access System [2]. On a higher abstraction level, we view the service as a provider of a continuous stream of news articles retrieved from news Web sites.

Classification Service: This Service was built on SVM^{ligth}, a package written for automatic text classification using support vector machines [5]. This enables the classification of the gathered news articles into a reduced, but uniform, set of categories based on train examples from pre-classified news. In our prototype the examples were taken from an existing publication that acted as our classifier reference [7].

Topic Collection Index Generators: Together the two components presented above load a common database with the indexed and classified news articles. From this database, specific topic collection index generators can be built to provide updates to topical Web portals (also known as vertical portals).

The remainder of this section presents the components of our news information management system in detail.

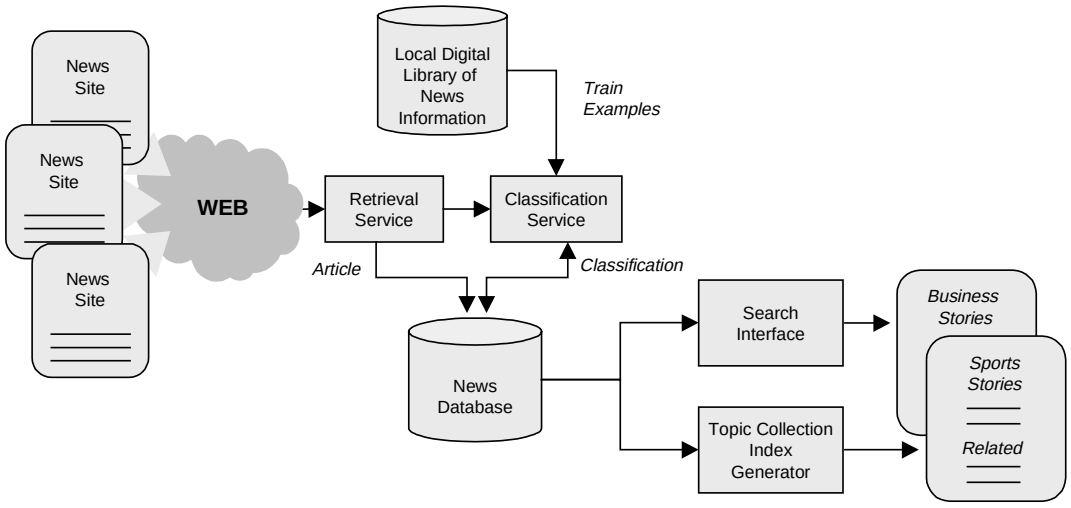

Figure 2. Architecture for retrieval and classification of news articles. The Retrieval Service collects articles from news sites distributed over the Web. These articles are sent to the Classification Service for classification on a common scheme according to train examples provided by a Local Digital Library. Classified articles are then stored in a News Database, ready to use by custom publications. Topic Collection Index generators update vertical Web portals.

3.1 Retrieval Service

Figure 3 details the Retrieval Service behavior. Multiple gathering agents retrieve and index news Web sites. To deal with different periodicity of publications, several queues are created and managed expressing different update priorities and different gathering schedules (Q_{D1} for daily publications, Q_{W1} for weekly editions, etc.).

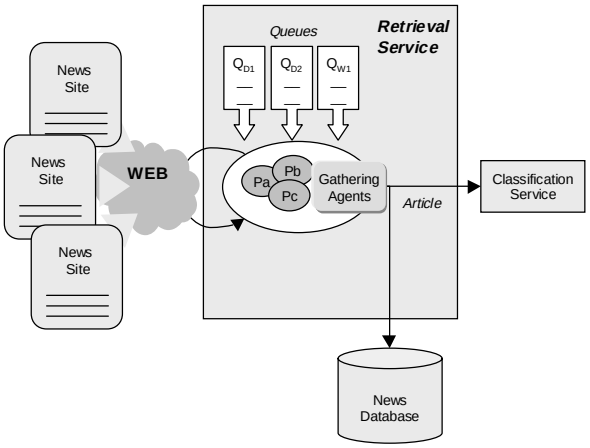

Figure 3. The Retrieval Service sub-components. Articles are retrieved and indexed from the Web by gathering agents, according to a pre-defined periodicity queue. New articles are stored in the Database.

In general, as we are only interested in indexing the current edition of each publication site, gathering agents can be configured to index only part of the available information in the site. This capacity allows the scalability of our solution, avoiding indexing or re-indexing of past news or irrelevant information, despite the fact that a simple header comparison in some cases would prevent re-indexing. However, in some cases, news articles are dynamic Web pages built on demand, which would invalidate any header information in our database.

3.2 Classification Service

Figure 4 details the Classification Service sub-components and their behavior. 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.

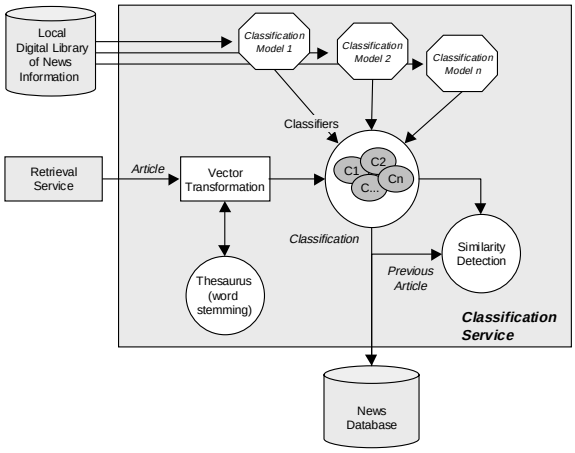

Figure 4. The Classification Service sub-components. Articles received from the Retrieval Service are converted to a vector format according to a predefined vocabulary. This enables classification based on the models defined for each category. Classification results are loaded in the News Database. Upon insertion, each new article is compared with a set of recent articles checking possible similarities.

Actual classification, based on pre-defined classification models, one for each 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. In our implementation this corpus is taken from a daily newspaper containing articles manually classified by their editors.

Once the Classifier processes one article, the resulting classification confidences are loaded in the News Database. We store the article's confidence level, returned from the classifier for each model, and information about possibly related articles. A Similarity Detection mechanism is run to quantify proximity between each new article and recently processed articles. If similarity is detected, then the article is grouped with other articles on a related cluster.

3.3 News Database

Figure 5 presents the data model for the news articles database. Each loaded article has the common attributes shown in the Figure 5. Some attributes of this set are extracted with heuristics, as the document structure of the target publication is unknown. This is the case of "article-bodytext." It is important that the article text, delivered to Classification Service, is clean from navigation bars or headers commonly found in the HTML of Web news information. These act as noise that can induce the classifiers in misclassifications. This is indeed data specific of each site, which does not necessarily represent the features that identify the themes.

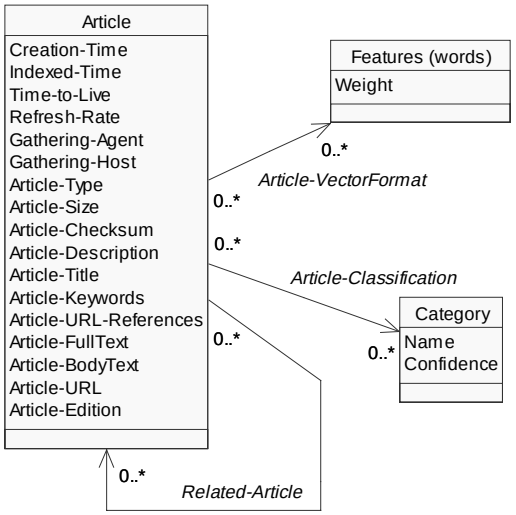

Figure 5. Data model for the article database. News articles are catalogued with the set of attributes shown. Its vector format is stored with the association to Features and the article's classification is stored with the confidence level for each category. Articles can also be related to other articles.

In our data model, we also store each article's vector format as associations to the Features class. The weights of the features (words) in a document are calculated with the Inverse Document Frequency - IDF based on the vocabulary defined by the trained examples. The weights are also normalized. This format is used in the Similarity Detection mechanism to determine possible relations with other articles through a similarity function. If similarity is detected, then an association with other articles is stored creating a cluster of related news.

The article classification activity is not exclusive: an article may belong to several categories. Each classification model expresses an absolute confidence value for the article. So, for each article, multiple associations with categories and confidence values are stored.

3.4 Theme-based Portals Generator

Figure 6 presents an overview of the access methods to the News Database.

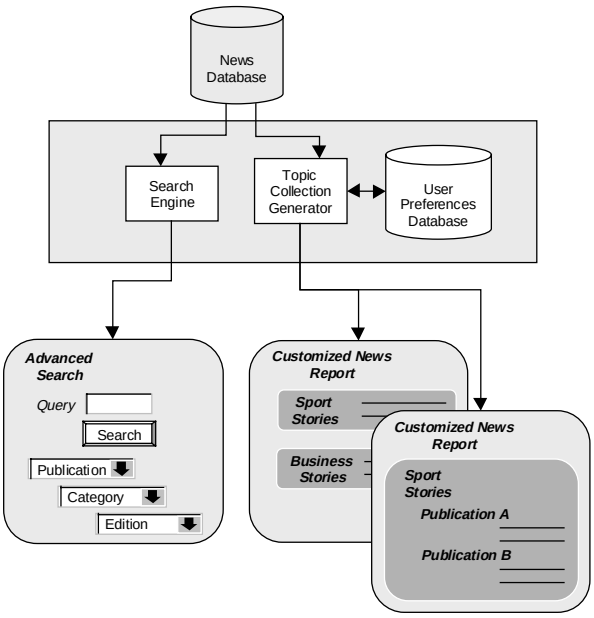

Figure 6. Interface to the News Database search engine. Users may submit advanced queries specifying multiple restrictions or access customized news reports, built according to stated personal preferences.

The database can be accessed through an advanced search interface where the user can customize the query with restrictions according to the data model. The interface enables search and retrieval of news articles from selected news Web sites filtered by category or edition. The entity that handles submitted queries is an application server that in turn uses the Glimpse search engine from the Harvest system [6]. This search engine uses a simple query format but handles complex and advanced queries efficiently and scales well.

The News Database can also be used to generate customized news reports. These are dynamically built with preferences supplied by the users. Preferences are stored in an autonomous database that associates each user with a list of rules. These define

the format for a personalized publication containing a list of potential highly relevant references to the user's information interest. The Topic Collection Generator also uses Glimpse as the search engine in the News Database.

3.5 Training Classifiers Strategy

Selection of the appropriate training information is crucial in our system, as it determines the classifiers' efficiency and accuracy.

The news articles used in training are previously filtered. Filtering consists in cleaning new Web pages from HTML tags and all other noise that they may have related to formatting and navigation.

The selected articles were picked from a reference newspaper and have publication dates uniformly distributed along the last year. We have observed that this approach minimizes the negative effects of cyclic events.

Our classifiers use the maximum number of features available, as the SVM-based classification mechanism scales well and we do not suffer the performance degradation of traditional classification algorithms.

4. RESULTS

The developed theme-based news management system provides a framework for multi-category classifications. For each classification model tested, if the confidence value returned by each model is above a defined positive threshold, we classify the article in the corresponding category. We validated the classification models from our local library against other publications. These achieved good accuracy, 94,5% with non-exclusive classification and using contents filtering. This precision was accomplished with a sample of articles from six publications representing one day of news (approximately one thousand articles).

We tested classification on full-text articles, as they appear on the publishers' sites (N/Filt.), and with contents filtering (Filt.). The gap between each method precision is about 5% (only 89,8% correct classifications against 94.5% with filtering strategies). Figure 7 presents the results for a set of categories on classification confidence using these two different methods. The confidence value is the absolute classification score representing the distance between the hyperplane dividing positive from negative examples and the point represented by the article (regarding SVM theory). Confidence is ultimately reflected in classification precision.

Filtering of format elements and navigation bars, generally present in Web documents, is performed by applying a set of heuristics specific of scanned news sites. Degradation in classification confidence of unfiltered news is more evident in general themes such as 'Society', 'Politics' or 'Culture'. However, in more specific categories, like 'Sports', 'Business' or 'Regional', accuracy does not improve significantly with filtering.

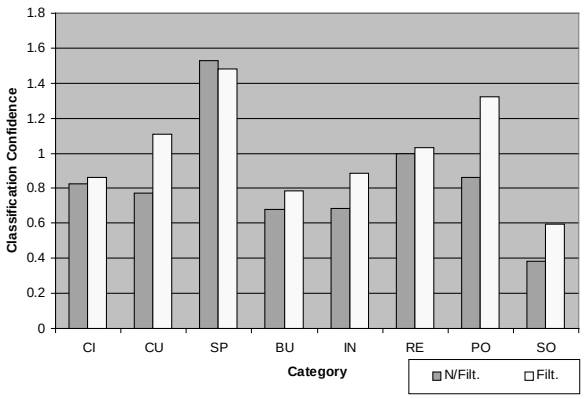

Figure 7. Comparison of the confidence in classifications on test examples with a limited set of categories. (Categories: CI: Science, CU: Culture, SP: Sports, BU: Business, IN: International Politics, RE: Regional News, PO: Politics, SO: Society)

In the extremely dynamic environment of Web news we also must be aware of the degradation of the classifiers' accuracy in time. Figures 8a e 8b 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.

With the first strategy, the efficiency of the classifier decreased tracking seasonality of the news events. This behavior is more apparent in dynamic categories such as 'Sports' or 'Politics'. In more static domains, such as 'Culture' or 'Business', classifiers have shown to be more stable.

The decreasing efficiency tendency is not visible using the second strategy, where classifiers, built with a higher number of articles spread along the year, present more stable levels of accuracy.

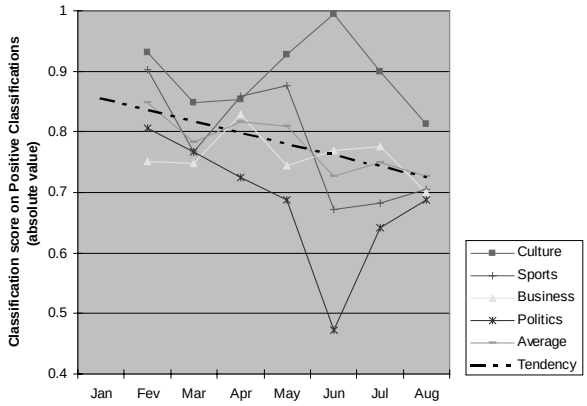

Figure 8a. The classification score on positive examples in time, obtained with models trained with articles from January. In topics with low seasonality such as culture, confidence is higher, but in other topics, like politics, confidence may decrease sharply.

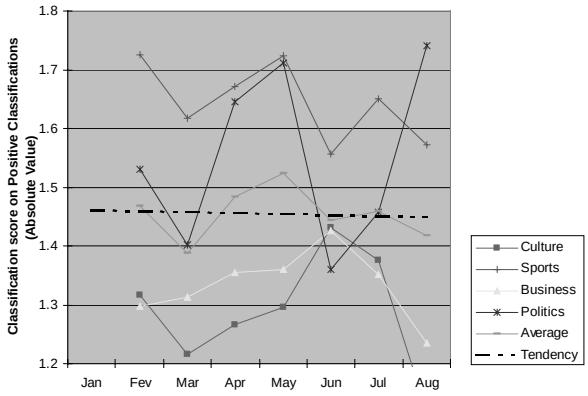

Figure 8b. The classification score on positive examples in time, obtained with models trained with articles of a full year. Confidence in classifications is significantly higher than with the previous strategy. Although some variations are visible the overall behavior is more stable in time.

The degradation of accuracy, measured by the confidence on positive classification, obtained by linear regression of the average behavior with the first strategy is however, minimized with our classifier building strategies, which do not force any feature selection. On the other hand, the above results have been obtained for a reduced number of wide scope categories. For a more specialized collection of categories we can expect a sharper degradation.

We are now finishing the similarity detection mechanism. Once completed, we will measure the accuracy of the used clustering techniques.

5. CONCLUSIONS

The increasing number of news Web sites on the Internet today demands specialized services to efficiently manage relevant news. We proposed an architecture to gather, retrieve and classify news articles and a model to manage this data.

Our classifiers, based on Support Vector Machines, presented good performance when applied to the categorization of news (accuracy of approximately 94%), confirming the ability of SVM's to build accurate and efficient classifiers. To accomplished this result we preprocess each article, filtering presentation noise with specific heuristics. We estimated in 5% the increase in accuracy when filtering is applied.

Our experiments have also shown that for news classification, accuracy degradation in time is minimal for some themes, but can be very high when seasonality is present. This can be minimized with criterious selection of train examples, sampled from a long time interval to compensate for the seasonality of the topics and periodic retraining of the classifiers.

The news classification system, discussed in this paper, is now being submitted to user evaluation. We expect to have user feedback regarding its usefulness soon.

6. 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] 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.

[4] Joachims, T., Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the Tenth European Conference on Machine Learning* - ECML, 1998.

[5] 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.

[6] Manber, U., and Wu, S.. Glimpse: a tool to search through entire file systems. In *Proceedings of the USENIX Winter Conference*, pages 23-32, 1994.

[7] 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.

[8] Vapnik, V.N.. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[9] 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.

[10] 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.