Personal News RSS Feeds Generation Using Existing News Feeds

Bin Liu, Hao Han, Tomoya Noro, and Takehiro Tokuda

Department of Computer Science, Tokyo Institute of Technology
Meguro, Tokyo 152-8552, Japan
{ryuu,han,noro,tokuda}@tt.cs.titech.ac.jp

Abstract. Nowadays more and more news sites publish news stories using news RSS feeds for easier access and subscription on the Web. Generally, news stories are grouped by several categories and each category corresponds to one news RSS feed. However there are no uniform standards for categorization. Each news site has its own way of categorization for grouping news stories. These dissimilar categorization can not always satisfy every individual user, and generally the provided categories are not detailed enough for personal using.

In this paper, we proposed a method for users to create customizable personal news RSS feeds using existing ones. We implemented a news directory system(NDS) which can retrieve news stories by RSS feeds and classify them. Using this system, we can recategorize news stories from original RSS feeds, or subdivide one RSS feed to a more detailed level. With the classification information for each news article, we offer customizable personal news RSS feeds to subscribers.

1 Introduction

At present, there are lots of news sites on the Web. Many of them offer news RSS feeds¹ for easier access and subscription. News RSS (Really Simple Syndication) feed is an XML-based format document for sharing and publishing frequently updated Web news. By subscribing to some news RSS feeds using a RSS reader, we can get alerts about publications of new issues. Generally news sites divide news articles to numbers of categories and publish news RSS feeds corresponding with these categories one-to-one.

Unfortunately, there are no uniform standards for categorization, news sites determine how to categorize news articles by themselves. For example, *CNN.com*² provides news RSS feeds by fields such as *Science*, *Sports*, *Business* and so on, while *allAfrica.com*³ offers news RSS feeds grouped by countries or regions. As we can see, there are some difference of categorization between various sites. If users happen to find just what they want in the given categories, the categorization is

---

¹ <http://cyber.law.harvard.edu/rss/rss.html>

² <http://www.cnn.com/>

³ <http://allafrica.com>

contributing. While if users can not find any appropriate categories close to what they want, the categorization does not make any sense. For instance, if users want to subscribe to news about diseases from *allAfrica.com*, they have to subscribe to all of the news RSS feeds from this site and discriminate the news about diseases one by one by themselves. So the original categorization of each site can not always satisfy every individual user. Further, usually categories used in news sites are not subdivided. They are not detailed enough for personal using. This make users have to handpick what they really need from the news gotten from RSS feeds.

Some RSS reader tools can let subscribers integrate RSS feeds, however what these tools could do is only to make a union from selected feeds by users, they do not make any analysis about the contents in feeds. News alert can also filter useful news stories for users, while users have to imagine all the presumable expressions for keywords and connect them together with *OR* during the initial setting of alerts. It is acceptable when the keywords are technical terms, but we could know what we omit when the keywords are general words. Further, simple string matching is used in news alerts, they will give a hit when *ice hockey* occurs while user wants stories of *hockey*.

In this paper, we propose a method for recategorizing the articles published from existing news RSS feeds, and using these subdivided news articles, we provide personal news RSS feeds for users. Personal news RSS feeds can be configured for individual demands like Fig.1. Users can recategorize or subdivide the news articles gotten from existing news RSS feeds according to their individual

Fig. 1. Overview of personal news RSS feeds

needs. We implemented a news directory system which gives preconditions for re-categorization and subdivision. It retrieves news articles using information from existing news RSS feeds and subdivides news articles into categories automatically. We gave each category used in this system a definition, and constructed automata with these definitions to categorize news articles with high speed. Each definition includes several related expressions (synonyms and abbreviations) of the corresponding category. So users need not make an association of all the expressions of their interested topics. We also avoid miss hits like *hockey* and *ice hockey* using limitation in categories' definitions.

The organization of the rest of this paper is as follows. Section 2 gives an overview of our news directory system. Section 3 presents the mechanism of automatic retrieval and subdivision of news articles, the basis of our work. While section 4 shows directions about how to get personal news RSS feeds using existing ones. Experimental results demonstrating the effectiveness of our approach are in Section 5. Section 6 discusses related work. Finally conclusions and directions for future work can be found in Section 7.

2 Overview of NDS

News directory system can be divided into two subsystems as Fig.2. One is for news retrieval, the other is for classification. System for news retrieval detects news titles and news body from original pages using the information of news titles and URLs. And the system for classification categorizes news stories with automata constructed from definitions of categories. We can get the results of categorization by scanning news stories only one time. We will give introduction about these two subsystems respectively in following section.

Fig. 2. Structure of News Directory System

3 News Directory System

3.1 Automatic News Collection

In this section, we give a brief introduction about the process of automatic news collection. Detailed explanation can be found in paper [11]. As a general approach, pattern matching is used in extraction from web pages. Considering it need corresponding patterns for various web site, extensibility is low when we get new web sites. We collect news articles by extracting news titles and body from news pages using information in original news RSS feeds.

The initials “RSS” are used to refer to the following formats:

- Really Simple Syndication (RSS 2.0)
- RDF Site Summary (RSS 1.0 and RSS 0.90)
- Rich Site Summary (RSS 0.91)

Although there are a number of different formats of RSS, all of them include the URL and title information in <link> and <title> respectively. These two information fields are the minimum necessary parts of each news item in a RSS feed. We detect these information and extract news articles from original news pages. The phase of the news article extraction consists of the following two parts.

Detection of News Title. The process detects position of the news title in the original news pages. Since the title shown in news feeds is not always same as the real title in original news pages, we have to try to extract news titles from original news pages once again. And because of the difference between titles in RSS feeds and original news pages, exact match is not appropriate for news titles detection. Instead, for each node n in the news pages (an HTML document⁴), we calculate similarity score with the news title in RSS feeds.

If the score is higher than a predetermined threshold, the string covered by the node n is judged as a news title. If there is no node whose score is higher than the threshold, no string is judged as the news title. On the other hand, if there are more than one node with higher score than the threshold, all of the strings covered by the nodes are judged as news titles.

Extraction of Body of News Articles. The process detects a part of the news article body and extract the whole body. Since body of a news article is usually preceded by its title, the process tries to find the news article body in some “contents ranges” at first, and, if it cannot find out the body in the range, it tries to find the body in a “reserve range”. “Contents range” and “reserve range” are parts which might include the news article body. They are determined as follows.

- If only one string is judged as a news title in the previous process, the following part and the preceding part are a contents range and a reserve range respectively (Fig.3(a)).

---

⁴ <http://www.w3.org/TR/html401/>

Fig. 3. Contents range and reserve range

- If no string is judged as a news title, the whole part of the news article page is a contents range and no reserve range exists (Fig.3(b)).
- If more than one string are judged as news titles, for each of the strings except the last string, range of between itself and the next string is a contents range. The part preceded by the last string is also a contents range. The part followed by the first string is a reserve range (Fig.3(c)).

At first, we specify a part of news article body. Then we calculate possibility score of each leaf node with non-link text n in each of the contents ranges. If there were some nodes with higher score than a predetermined threshold, we consider the nodes with the highest score cover a part of the news article body. Otherwise, we consider the nodes with the highest score in the reserve range cover a part of the news article body. Since a news article body is usually a continuous text, it can be extracted by taking leaf nodes around the specified nodes. However, in some cases, some information which is not related to the article, such as advertisement, is inserted in the article body. In order to avoid taking such information, we also set limitations to filter them. Finally, we get a list of nodes which cover the whole news article body. The whole body can be extracted by getting the node value (i.e. text) from each node in the list.

3.2 Automatic News Classification

After the news articles extraction, we get the materials for news classification. The next step is to give categories for classification and define them to construct automata.

News Categories. At first we need categories for classification. In news directory systems we use one-level flat directory structures or multi-level tree directory structures. Typical examples of one-level flat directory structures may be as follows.

- Classification of natural disasters such as typhoon and earthquake.
- Classification of human diseases such as diabetes and malaria.

Fig. 4. Composition of a small classification tree

Fig. 5. category *Disease*

Fig. 6. category *Countries/Regions*

Typical examples of multi-level tree directory structures may be as follows.

- Classification of locations such as countries/regions on the earth and outside of the earth.
- A small classification tree constructed from the large classification tree such as *WordNet*⁵ [9] or *Wikipedia*⁶ structures.

An example of one-level flat directory structure is shown in Fig.5 and an example of multi-level tree directory structure is shown in Fig.6. Users can also build their original directory structures manually. Here we give methods to build directory structures with existing resources.

⁵ <http://wordnet.princeton.edu/>
⁶ http://en.wikipedia.org/wiki/Main_Page

Method 1. We use open knowledge collection of classifications by humans, such as *Wikipedia* and *WordNet*, to build an initial collection of instance names belonging to one category.

Method 2. Our method of building multi-level tree directory is as follows. We need a small set of basic words. Such a set of basic words may be subject words in *New York Times Topics Index*⁷ or a subset of *Longman defining vocabulary*[13] or a subset of *Oxford defining vocabulary*[3]. For a given set of basic words we construct a small classification tree as follows.

1. We retrieve full paths of all basic words in the *WordNet* tree.
2. We construct the initial small tree using the full paths obtained in the step 1.
3. We construct the small tree by deleting all non-basic words having exactly one child node from the initial small tree.

A process of construction of a multi-level tree directory is shown in Fig.4.

Automatic Placement. In order to realize the automatic placement, each category need a definition. Our default definition of a news article *A* to be contained in a category *B* is that the article *A* has an occurrence of the word *B*. In addition to default definitions of single word occurrences, we use explicit definitions of a news article in a category using the expressions defined by following extended context-free syntax rules with repetition operator $\{ \}$ representing zero or more times of repetitions.

$$\begin{aligned}
 expression &\rightarrow (term) \{OR (term)\} \\
 term &\rightarrow factor \{AND factor\} \\
 factor &\rightarrow (phrase)|(NOT phrase) \\
 phrase &\rightarrow word \{SPACE word\} \\
 word &\rightarrow character \{character\}
 \end{aligned}$$

This expression allows us to define news articles having slightly more complicated word occurrences. For example, we may write a definition for category *soccer* using the following expression.

$$((\text{football})\text{AND}(\text{NOT american\_football}))\text{OR}((\text{soccer}))$$

This expression means that an article *A* is to be contained in the category, if *A* contains the word *football* but not *american football* or *A* contains the word *soccer*. The same expression may be written briefly as follows.

1. football **AND** (**NOT** american football)
2. soccer

We collect phrases from dictionaries of synonyms and append the **NOT** limitation according to the inclusion relations among the phrases we used. And then

⁷ <http://topics.nytimes.com/top/reference/timestopics/>

Fig. 7. M_1

Fig. 8. M_2

give **AND** limitations where **NOT** appears to create terms. At last connect all the terms for same meaning with **OR** limitation. Using the definitions, we make keywords matching to realize automatic placement. Making simple comparison between target string and source string costs much time. We realize this process more efficiently by using finite-state automata, which allow us to get the results of classification by scanning the news articles only one time.

The task of automatic placement consists of two phases using finite-state automata. In the first phase, we construct an automaton with all the phrases used in category definitions, it can help us to detect which phrases we used in definitions appeared in the news story. And in the second phase, we construct another automaton with all the limitations used in definitions, it can tell us which definitions the new story satisfied. We call these two automata as M_1 and M_2 . For the sample expressions of category *soccer*, we can construct M_1 and M_2 shown in Fig.7 and Fig.8. About the details of M_1 and M_2 , we introduced in paper [17].

4 Personal News RSS Feeds Generation

After the news extraction and classification, we can use the results of news' classification to help users generate their personal news RSS feeds. We explain the process of personal news RSS feeds generation in this section.

At first, news sites or news feeds should be designated for contents extraction. We offer users about 40 well-known news sites such as *CNN*, *BBC*⁸ and so on, and RSS feeds from these sites. While we do not mean to put restrictions on users' sites selection. Users can keep their favorite news sites or news feeds as usual. If only the users favorite news sites publish RSS feeds, and they could designate the URLs of RSS feeds. Then our system will also operate extraction and classification.

Secondly, user can select the categorization or categories which they are interested in. We provide categorization such as *countries/regions*, *human/organizations*, *events/accidents*, and so on. Each categorization has numbers of categories which may have a tree structure. If users could not find a appropriate categorization or categories. They can also input the keywords for filtering

⁸ <http://www.bbc.co.uk>

certain topics. In this case, our system will create a personal automaton for classification using the input keywords.

Personal news RSS feeds will be helpful in following two cases.

1. Replace the categorization of original RSS feeds.

If users wanted to read news articles grouped by countries or regions from a news site which only provides news feeds in categories like *Science*, *Sports*, *Business* and so on. Users can designate the URLs of original RSS feeds and subscribe to the categorization of countries or regions. Contents would be sent to users in several RSS feeds and each feed corresponds to a country or region. User can also subscribe to news feeds of certain countries or regions by designating certain categories in the categorization.

2. Subdivide the news of original RSS feeds.

User can subdivide the news in RSS feeds by operating categories. For example, we can get a news feed which sends news articles about both *whale* and *Japan* by making a intersection set of categories *whale* and *Japan*. The order of intersection will result in different meaning. If we selected *Japan* and then *energy*, we would get news articles grouped by kinds of energy, and all the news articles also belong to category *Japan*. If we selected *coal* and then *Asia*, we would get news articles grouped by countries or regions in Asia, and all the news articles also belong to *coal*.

According to the usages mentioned above, personal news RSS feeds are generated by following steps.

1. Pick up sites and news RSS feeds from the lists we offered. If users' favorite sites or RSS feeds were not in our lists, users can also register the URLs of the new RSS feeds into the system.
2. Select the categories or make intersection sets from the given categories.
3. New personal news RSS feeds are generated according to the results of user's choices. An unique URL is issued for the personal feed.

Once personal news feeds are generated successfully, users can register the feeds' URLs into their RSS reader tools. Our system will send along corresponding news articles to users by the personal news feeds at fixed intervals.

5 Experimental Results

In this section we introduce our implementation in details. We also evaluate our approach using the results of experiments.

5.1 Implementation

We implemented the parts of news articles extraction and classification. The news sites from which we collect news articles are 40 sites from 21 countries or regions and news RSS feeds from these sites are 624 in all. We run the extraction

Table 1. Result of News Extraction

1000 news pages		
successful extraction	extraction failure	
902	partially-extracted	non-extracted
	68	30

at fixed intervals and we can get about 1,500 latest news articles each time averagely. We constructed a directory structure for our news directory system using resources from Wikipedia and other existing resources. We also constructed a small classification tree of 885 nodes with 624 basic words from Longman defining vocabulary and 261 non-basic words from WordNet. Categories are given in methods like *Countries/Regions*, *Sports*, *Diseases* and so on. The max value of the depth in the directory structure is 5. The part of automatic classification is also implemented. We used 2,328 terms in all for the definitions of the 825 categories and automata M_1 and M_2 are generated with 12,801 and 1,666 nodes respectively.

5.2 Evaluation

Using the news directory system, we collect news articles and make them classified. We evaluate our approach and system in following sections.

Automatic News Collection. We selected 1,000 news articles from the results of extraction in random order and compared them with the original news contents in each corresponding news page. Results is shown in Table 1. We found 970 articles were extracted successfully and most of the cases of failure are due to multi-pages, that is, when the contents of a news article is too long to show in one page, most sites will divide the contents into several parts and prepare one Web page for each part. In this case, our approach just extracts the partial contents on the first page. We can also find some advertisement, blog pages or video news in the RSS feeds of some news Web sites, and news articles in some news Web pages can not be viewed until users log into the news sites. Our approach can not extract well from these Web pages.

Automatic News Classification. We manually evaluated the precision rate and recall rate of our automatic classification method using *country/region* classification of 500 news articles, the results are shown in Table 2. In these

Table 2. Result of Automatic Classification

500 articles		
articles classified appropriately	inappropriate articles	
453	not classified	misclassified
	12	35

Table 3. News count from feeds of *sports*

Name of feed	news count	Name of feed	news count	Name of feed	news count
Sports	219	Golf	28	Archery	55
Athletics	7	Baseball	4	Basketball	1
Boxing	9	Cycling	4	Diving	1
Fencing	1	Gymnastics	1	Hockey	11
Rowing	1	Sailing	1	Swimming	10
Weightlifting	1				

500 news articles, 453 articles are appropriately classified. 12 articles mentioning country/region names are not classified into any category of country/region, because our definitions of corresponding categories did not contain the expressions used in those articles. 35 articles not mentioning country/region names are classified into countries/regions, because company names, event names, and news source names may contain country/region names. Because we did not use semantic analysis, system can not pick out multisense words yet at the present time.

Personal News RSS Feeds Generation. We supposed a user wanted to subscribe news stories about sports from *BBC*. Because there is no RSS feed corresponding to sports from *BBC*, user has to input all the keywords related to sports to set up a news alert. Instead, when we customize a personal RSS feed from all *BBC* feeds with category *Sports*, what we need to do is only to check some checkboxes. We checked this personal feed from Jan, 2009 to Feb, 2009, 219 stories were sent to our RSS reader in 16 kinds of sports as Table 3, and we made a search with keyword *sport* in news articles from *BBC* in this period, there were only 19 hits. We also took same experiments at other sites.

6 Related Work

Our approach contains news extraction and automated classification. So we will mention related work about these two topic respectively and give comparison with other systems.

6.1 News Extraction

There are two opposite approaches to the recognition and extraction problem:

1. Static patterns

In this approach static patterns (extraction template) need to be defined previously for every source indexed by the system. Each web site has its own source structure of pages and the document location would be different, too. So in the extraction phase the pages of every site are individually processed filtering the documents.

The advantage of this kind of methods is the computational cost. On the other hand, a lot of human intervention is needed. For every new source to

be added to the system, users should analyze the internal HTML structure of the documents and define a custom template. If some site changed something in publication format or the document structure, the corresponding template must be redefined. Thus the system maintenance becomes in a critical task.

2. Automated extraction

Most of the published works belong to this approach. These techniques aim to avoid the human intervention and enable dynamically source adding to the systems. There are mainly two ways to affront the automated solution:

- Adaptation of data extraction

Traditional techniques based on different clustering techniques as for example tree edit distance [7,18], or use of equivalence classes [2]. The concept over these approaches lie, is that news with common structures will match in the same cluster or class, so after the clustering phase a extraction template could be generated for each cluster. This implies multiple reprocessing of the documents with prohibitive computational cost. Thus this family of techniques is not applicable in real systems, it is only useful in applications where the number of documents managed is reduced and the frequency of content update is low.

- Domain specific approaches

Other approaches try to combine the previous knowledge in the area of data extraction taking in account the singular characteristics of the news domain. Some works try to exploit the structure of the articles by semantic partitioning [16]. This approach is not still computational efficient and the results of precision and recall claimed by the authors can be improved. Other recent work [9] tries to use the tables present in the documents after assume that the news are present in the larger cell. Actually this assumption is false in most of the cases the news articles are not contained by tables. Also the evaluation methodology used in this work is very poor.

So in this context we present an automated extraction approach based on the provided RSS feeds. With the information of news title and URL, we detect news contents from the original pages. Our method is a tradeoff between computational efficiency and result effectiveness.

6.2 Automated Classification

Automated classification is also a well studied problem. There are two main approaches to realize automatic text classification.

1. Clustering Clustering [14] is a common technique to divide objects into several groups (called clusters). Objects from the same cluster are more similar to each other than objects from different clusters. Usually similarity is assessed according to a distance measure. There were also some experiments [4] taken to apply clustering to classify news articles. Well, this application showed us which news groups (cluster) will occur after analysis. It is unsuitable when users know definitely what kind of news they want.

2. Classification

Classification is distinguished from clustering by whether there are categories given previously before processing. The following two kinds of approaches are mainly applied to realize classification.

- Hand-Crafted Rules

*Google alerts*⁹ takes this approach to filter information for users. It needs users to give a set of keywords which they think are important to set up. If the occurrences of these keywords were detected, system will notify users.

The advantage of this approach is that rules can be created simply by listing related words. By the same token, system could only detect the words listed because of the exact matching [1]. In the same way, system will tell us there is a hit when it detects *ice hockey* even we adopted *hockey*. While, when there are numbers of categories, we have to define them one by one, too. So we cannot use this approach directly.

- Machine learning

Machine learning has demonstrated good performance can be achieved on spam/junk email. For example, *SpamCop*[12] (Pantel & Lin, 1998), using a Naive Bayes approach achieved accuracy of 94%. Sahami [15] applied a Bayesian approach and achieved precision of 97.1% on junk and 87.7% on legitimate mail and recall of 94.3% on junk and 93.4% on legitimate mail. Besides approach of Bayes, TF-IDF [5], K-Nearest Neighbor [19] and SVM [6] are also common applied techniques. Well, using machine learning to classify news article, we need numbers of labeled documents to create a model at first. Labeling must be done by a person, this is a painfully time-consuming process and it is per se impractical for news categories which are changing momentarily. No one would like to be ordered to gather numbers of samples when he (or she) plan to create a new category.

In this context, we use the rule-based method and proposed automated method to construct categories and rules (definitions). And we use limitation in definitions to avoid miss matching like *ice hockey* and *hockey*.

6.3 Comparison

Comparing with *Google Alerts*, user do not need list all the expressions of a topic they are interested in with our approach. Because we have considered most of the possibilities of expressions about a category during the process of defining categories already. So the necessary operations become more simple and the recall rate of our approach is higher than that of *Google Alerts*. And another thing, because simple string matching is used in *Google Alerts*, when keyword A is contained by keyword B completely (such as *hockey* and *ice hockey*), there may be some mistakes in the results if users input keyword A. In our approach, we avoid this kind of mistake by using **NOT** relation in categories' definitions.

---

⁹ <http://www.google.com/alert>

*NewsKnowledge.com*¹⁰ provides a more friendly service. This site allows users create personal news RSS feeds. Categories are subdivided and users can choose their favorite categorization such as *Health*, *Industries*, and so on. Users can also give keywords for filtering certain topics. However, the source of news feeds are limited, so users can not designate their favorite news feeds or news sites. And the subdivided categories are still in an insufficient degree.

7 Conclusion

In this paper, we have presented an approach for generating personal news RSS feeds from existing news feeds using news extraction and automatic classification. We also proposed methods to realize the news extraction and automatic classification. We implemented the methods and confirmed the availabilities of our approach.

As our future work, we will try the news articles extraction from multi-pages, enrich the news sites and news feeds, the categories in news directory, and improve the precision rate by resolving the problem of multisense words, too. We also plan to develop a RSS reader tool which allow users view news feeds multi-level structure, that is, users can view parts of the directory structure of our news directory. Users can view the other items in the category which contains the item they chose, this could be suggestive for users.

References

1. Aho, A.V., Corasick, M.J.: Efficient string matching: an aid to bibliographic search. *CACM* 18(6), 333–340 (1975)
2. Arasu, A., Garcia-Molina, H., University, S.: Extracting structured data from web pages. In: *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pp. 337–348. ACM Press, New York (2003)
3. Hornby, A.S., Ashby, M.: *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford (2005)
4. Das, A., Datar, M., Garg, A.: Google News Personalization: Scalable Online Collaborative Filtering. In: *Proceedings of the 16th international conference on World Wide Web*. ACM Press, New York (2007)
5. Boone, G.: Concept features in re:agent, an intelligent email agent. In: *Second International Conference on Autonomous Agents*
6. Brutlag, J., Meek, C.: Challenges of the email domain for text classification. In: *Seventeenth International Conference on Machine Learning*
7. Reis, D.C., Golgher, P.B., Silva, A.S., Laender, A.F.: Automatic web news extraction using tree edit distance. In: *Proceedings of the 13th international conference on World Wide Web*, pp. 502–511. ACM Press, New York (2004)
8. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29, 103–137 (1997)
9. Zhang, D., Simoff, S.J.: Informing the curious negotiator: Automatic news extraction from the internet. In: Williams, G.J., Simoff, S.J. (eds.) *Data Mining*. LNCS, vol. 3755, pp. 176–191. Springer, Heidelberg (2006)

¹⁰ <http://www.newsknowledge.com/home.html/>

10. Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J.: Indexing with WordNet Synsets Can Improve Text Retrieval. In: Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal (1998)
11. Han, H., Tokuda, T.: Web News Contents Extraction Using RSS Feeds. In: The Proceeding of Annual Meeting of Japan Society for Software Science and Technology (2007)
12. Pantel, P., Lin, D.: Spamcop: A spam classification & organization program. In: Proceeding of AAAI 1998 Workshop on Learning for Text Categorization, pp. 95–98 (1998)
13. Proctor, P.: Longman Dictionary of Contemporary English. Longman (2005)
14. Berkhin, P.: Survey of Clustering Data Mining Techniques Accrue Software (2002)
15. Sahami, M., Dumais, S., Heckerman, D., Horvits, E.: A bayesian approach to filtering junk e-mail. In: AAAI 1998 Workshop on Learning for Text Categorization (1998)
16. Vadrevu, S., Nagarajan, S., Gelgi, F., Davulcu, H.: Automated metadata and instance extraction from news web sites. In: WI 2005: Proceedings of the The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005), Washington, DC, USA, pp. 38–41. IEEE Computer Society, Los Alamitos (2005)
17. Noro, T., Liu, B., Van Hai, P., Tokuda, T.: Towards automatic construction of news directory systems. In: The 17th European-Japanese Conference on Information Modelling and Knowledge Bases, pp. 211–220 (2007)
18. Crescenzi, V., Mecca, G.: Automatic information extraction from large websites. *J. ACM* 51(5), 731–779 (2004)
19. Yang, S., Jian, H., Ding, Z., Hongyuan, Z., Giles, C.: IKNN: Informative K-Nearest Neighbor Pattern Classification. *Practice of Knowledge Discovery in Databases*, 248–264 (2007)