Fair News Reader: Recommending News Articles with Different Sentiments Based on User Preference

Yukiko Kawai[1], Tadahiko Kumamoto[2], and Katsumi Tanaka[3]

[1] Undergraduate School of Science, Kyoto Sangyo University
Motoyama, Kamigamo, Kita-Ku, Kyoto-City 603-8555, Japan
Tel.: +81-75-705-2958; Fax: +81-75-705-1495
kawai@cc.kyoto-su.ac.jp
[2] Faculty of Information and Computer Science, Chiba Institute of Technology
2-17-1, Tsudanuma, Narashino, Chiba 275-0016, Japan
kumamoto@net.it-chiba.ac.jp
[3] Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
Tel.: +81-75-753-5979; Fax: +81-75-753-4957
ktanaka@i.kyoto-u.ac.jp

Abstract. We have developed a news portal site called Fair News Reader (FNR) that recommends news articles with different sentiments for a user in each of the topics in which the user is interested. FNR can detect various sentiments of news articles, and determine the sentimetal preferences of a user based on the sentiments of previously read articles by the user. While there are many news portal sites on the Web, such as GoogleNews, Yahoo!, and MSN News, they can not recommend and present news articles based on the sentiments they are likely to create since they simply select articles based on whether they contain user-specified keywords. FNR collects and recommends news articles based on the topics in which the user is interested and the sentiments the articles are likely to create. Eight of the sentiments each article is likely to create are represented by an "article vector" with four elements. Each element corresponds to a measure consisting of two symmetrical sentiments. The sentiments of the articles previously read with respect to a topic are then extracted and represented as a "user vector". Finally, based on a comparison between the user and article vectors in each topic, FNR recommends articles that have symmetric sentiments against the sentiments of read articles by the user for fair reading about the topic. Evaluation of FNR using two experiments showed that the user vectors can be determined by FNR based on the sentiments of the read articles about a topic and that it can provide a unique interface with categories containing the recommended articles.

1 Introduction

As the amount of Web content continues to increase, users are more strongly demanding novel Web sites that provide better quality content. Web portal sites

B. Apolloni et al. (Eds.): KES 2007/WIRN 2007, Part I, LNAI 4692, pp. 612–622, 2007.

now gather high-quality content from many Web sites and provide integrated pages. Users can then access various kinds of information from these integrated pages without accessing many Web sites. With a news portal site on the Web such as GoogleNews and Yahoo! , users search for articles of potential interest from many collected articles by using keywords or selecting a category. Retrieval using aspects other than keywords (e.g. sentimental aspect of articles) is not supported.

Conventional news portal sites have two basic ideas behind article search and classification: one is the frequency at which keywords occur in an article[1][2][3], and the other is the structure of the links between pages[4][5]. While the user may be able to browse articles of potential interest listed in the search results, if he or she wants to read articles based on their sentimental aspects, he or she has to judge the sentimental aspects based only on the contents of the search results. For example, a search using the keywords "Iraq" and "terror" with existing news portal sites will return articles about "Iraq" with such topics as "suicide bombing in Iraq", which are likely to create a sad sentiment. A system that can recommend, for example, articles that create a happy sentiment, such as "released hostage in Iraq", should thus be useful.

We have developed a novel news portal system called Fair News Reader (FNR) that recommends articles that have symmetric sentiments against the sentiments of read articles by a user for balanced reading about each of the topics in which the user is interested. When the number of news articles about a topic that create the same or similar sentiments for a user is small, the articles are defined as "non mainstream articles" for the user. When the number is large, the articles are defined as "mainstream articles". The new technical contributions of FNR are as follows:

- mining sentiments of a news article,
- determining sentimental preferences for a user based on his or her browsing history, and
- discovering non mainstream articles for the user based on sentiments of articles and his or her sentimental preferences.

FNR recommends non mainstream articles to users in the following way. It first estimates the eight sentiments each article is likely to create: happy, unhappy, acceptance, rejection, relaxation, strain, fear, and anger. "An article vector" is created from each article; it has four elements based on the eight sentiments, and each element has a value ranging from 0 to 1, and is calculated based on the sentiments of words in the article. For example, when an element has the sentiment of "happy $\Leftrightarrow$ unhappy" from 1 to 0, if many of the words in an article create a happy sentiment, this element represented by a happy sentiment to be generated from the article has a large value. Next, the sentiments the user has about the topic of articles previously read are extracted by the article vectors and represented as "a user vector". The user vector for each topic is created from the article vectors of read articles, and calculated using the standard deviation represented by the fluctuations of each element of the vector by using read articles. Finally, for balanced reading about a topic, FNR recommends

non mainstream articles that have symmetric sentiments against the sentiments of read articles by the user. To find such non mainstream articles for the topic, FNR compares the article vectors with the user vector, and selects the articles for which each element of article vector has symmetric values of the user vectors' elements. Naturally, the recommendation algorithm can change the selections from non mainstream to mainstream articles (enabling the user to read articles more in line with his or her sentiments).

2 Overview of FNR

The FNR is implemented on the system which we have developed called "My Portal Viewer (MPV)" [2], which gathers and integrates articles from many news sites based on the interests of the user. In this section, we introduce MPV and then describe the concept of FNR on MPV.

2.1 News Portal System (MPV)

MPV collects articles from various news sites by crawling through them, stores the articles in a database, integrates the content as needed, and presents them on one page, the "MPV page". This involves two unique concepts: the user's interests and knowledge based on his or her access history and the MPV page, which emulates the "look and feel" of the user's favorite news page, as shown in Fig.1.

The MPV layout mirrors that of the users' favorite news page, and parts of the original content are replaced by the integrated content. In the example shown in Fig.1, the user had specified the CNN top page as his or her favorite news page, and some of the original content have been replaced by integrated

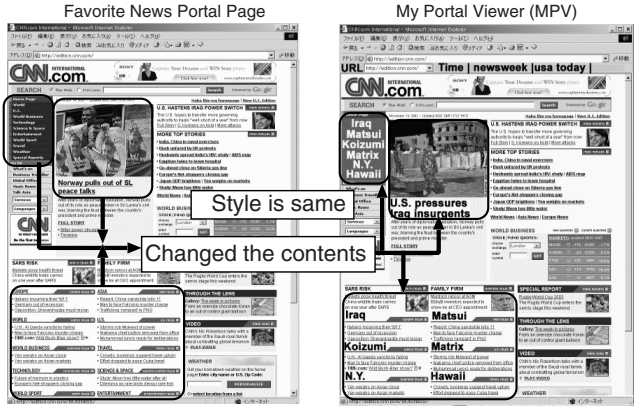

Fig. 1. Original information on favorite page is replaced with gathered information, and integrated information is shown on MPV page

content in the MPV page. Because the layout of the user's favorite news page is retained, he or she can easily locate particular information. The replaced content, categories, top news, and articles in each category, on the other hand, is created based on the frequency of term occurrence, as determined from the history of articles previously read by the user. The new category names are taken from the "interest keywords", which represent the topics in which a user is potentially interested based on the his or her access history. In the example shown in Fig.1, the categories "Iraq", "Koizumi", and "Matsui" were presented based on these keywords. Using the interest keywords as category names enables the user to easily grasp the content of articles in each category.

2.2 Concept of FNR

FNR uses the following three techniques developed for MPV. The first technique is that the gathered articles are categorized by interest keywords and "co-occurrence keywords", which are extracted from the user's browsing history. The second point is the category names are taken from the interest keywords. And the last point is that the layout of the user's favorite news portal page is used for the integrated portal page, and only part of the content is replaced by integrated content. For example, some category names are replaced with ones based on the interest keywords.

FNR offers the following three new techniques.

- It introduces new method of measure such as sentiments to news integration system.
- It recommends fair news articles for a user to him or her after it modeled the sentiments of the user about a topic.
- It integrates two or more categories that contain many of the same articles.

FNR is thus able to model a user's interests and sentiments, and the constructed model can be used to create a recommendation system.

The sentiments of the user about each article are represented by a vector with four elements, i.e. sentiments measure. They are based on Plutchik's eight basic emotions: joy, acceptance, fear, surprise, sadness, disgust, anger, and anticipation. Each scale represents two contrasting basic emotions; "happy ⇔ unhappy", "acceptance ⇔ rejection", "relaxation ⇔ strain", and "fear ⇔ anger". The value for each scale a real number between 1 and 0. For example, if an article has a value is 0.1 for "acceptance ⇔ rejection", the article should create a strong sentiment of "rejection". FNR calculates the average value and standard deviation for each scale. The four calculated standard deviations reflect the fluctuations in the interest keyword. If the standard deviation for a scale is larger than a threshold, FNR assumes that the user is not interested in news articles with a variety of sentiments in that category, and the value for that scale is defined as "*don't care*". If the standard deviation is smaller than the threshold, FNR assumes that the user is interested in news articles with biased sentiments in that category, and the value for that scale is set to the average value. As a result, the four scale have a value of "*don't care*" or the "average value", each of

which is determined by the standard deviation. The determined vector of the interest keyword and the interest keyword are defined as the user's preferences. FNR categorizes articles by the interest keyword, and it selects and recommends articles in each category by using this vector.

3 User's Sentimental Preferences for Fair Reading

In this section, we describe how the user's preferences are determined based on the user's interests and the sentiments for the user about previously read articles. FNR creates a user vector after extracting the interest keywords based on the user's browsing history.

3.1 Generating Vectors from News Articles

A user sentiment vector is generated for each input article.

1. Words whose parts of speech are action nouns, adjectives, or verbs are extracted from the information obtained from Web page P_i in step 1 for interest word extraction to be described in 3.2.
2. Scale value S_{je} and weight M_{je} for $e(e = 1, 2, 3, 4)$ of sentiment scale from 0 to 1 of each word j are obtained by consulting a sentiment dictionary, as described below.
3. Scale value O_{ie} for sentiment scale e of P_i is calculated:

$$O = \sum^{j} S_{je} \times |2S_{je} - 1| \times M_{je} \Big/ \sum^{j} |2S_{je} - 1| \times M_{je}, \tag{1}$$

   where the $|2S - 1|$ term denotes an inclined distribution depending on scale value S. When scale value S is 0.5, it is 0. When scale value S is 0 or 1, it is 1. Many of the words that appear in articles are independent of the feelings created by the articles. The inclined distribution described here was been introduced to remove the adverse effect such general words can cause in the calculations.
4. An sentiment vector for P_i is generated in the form of "$(O_{i1}, O_{i2}, O_{i3}, O_{i4})$".

The sentiment dictionary used in step 2 was automatically constructed by analyzing the Nikkei Newspaper Full Text Database[1] [6] using an extended version of the method proposed in Ref. [7]. The original method creates a sentiment scale from a pair of sentiment words, while our extended version creates an sentiment scale from two or more sentiment words. That is, we formulated which of two groups of sentiment words that composed an sentiment scale each of the words extracted from an input article would co-occur with more often

The groups of sentiment words used in constructing our sentiment dictionary are listed in Table 1, and part of the sentiment dictionary is shown in Table 2. The upper lines of each entry show the scale values, and the lower lines show the weights.

[1] This database has two million news articles accumulated over a 12-year period, from 1990 to 2001. Each edition consists of about 170,000 articles (about 200 MB).

Table 1. Impression scales designed for MPV Plus

Impression Scale	Impression Words
1. Approval – Disgust	Shonin (acceptance), Shonin-suru (approve), Aikou (love), Aikou-suru (love), Suki-da (like), Kyohi (rejection), Kyohi-suru (reject) Ken'o (aversion), Ken'o-suru (take an aversion), Kirai-da, (dislike)
2. Grad – Sad	Akarui (bright, encouraging), Ureshii (glad), Tanoshii (happy) Kurai (dark), Kanashii (sad), Kurushii (painful)
3. Relaxation – Strain	Yuttari (comfortable), Yuttari-suru (feel easy), Nonbiri (peaceful) Nonbiri-suru (feel relieved), Yukkuri (slowly), Yukkuri-suru (take one's time) Kincho (tension), Kincho-suru (become tense), Kinkyuu (emergency)
4. Anger – Fear	Okoru (get angry), Dogou (roar), Osoreru (dread), Kowai (scary), Kyofu (fear)

Table 2. Examples of entries in impression dictionary

Entry word	No. 1	No. 2	No. 3	No. 4	Entry word	No. 1	No. 2	No. 3	No. 4
Sosei (revival)	0.91	0.521	0.429	0.000	Shototsu-suru (collide)	0.344	0.353	0.315	0.529
	0.464	0.582	0.732	0.328		1.004	1.016	1.099	0.948
Shukkoku	0.596	0.209	0.762	0.201	Kenen-suru (worry)	0.373	0.319	0.246	0.293
(departure from a country)	0.975	1.049	1.065	0.701		1.447	1.440	1.521	1.275
Shibou (death)	0.28	0.358	0.260	0.364	Hofu-da (rich)	0.597	0.676	0.761	0.466
	1.132	1.272	1.306	1.112		1.416	1.352	1.299	1.109
Dassen (derailment)	0.31	0.546	0.403	0.291	Saiteki-da (optimum)	0.622	0.671	0.743	0.192
	0.514	0.603	0.737	0.549		1.185	1.164	1.145	0.899
Dekakeru (go out)	0.639	0.754	0.887	0.590	Konnan-da (difficult)	0.318	0.305	0.307	0.317
	1.430	1.394	1.304	1.114		1.451	1.526	1.528	1.274
Chosen-suru (challenge)	0.618	0.687	0.752	0.500	Fumei-da (unknown)	0.359	0.367	0.336	0.359
	1.399	1.330	1.251	1.090		1.241	1.337	1.364	1.18

3.2 Extracting Interest Keywords

1. FNR extracts and stores the metadata (title, description, URL, etc.) after it downloads pages $P_1 to P_n$ from several news sites.
2. The description and title are morphologically analyzed, and the proper nouns and general nouns are extracted.
3. The weight of each word is calculated using the term frequency and weight of the three parts of speech as in the following equation: $w_{ij} = tf \cdot idf = (log(F_j + 1)/log(F_{all}))\ (cdotlog(N/N_j))$, where F_j is the frequency of the appearance of word j in page P_i, and F_j is the frequency of appearance of all words in P_i. N is the number of all pages gathered, and N_j is the number of pages with appearance of a word j.
4. When a user reads articles on M pages, the weight W_j of word j on M pages is the summation of w_{ij}: $W_j = \sum_{i=1}^{M} w_{ij}$.
5. If W_j is larger than a certain threshold, j is identified as an interest keyword.

The detected interest keywords are used as new category names. FNR does not replace all the original category names because the number of categories on the original news portal page is limited. Instead, an "others" category is created, and the remaining interest keywords are placed there. When the user selects this category, the remaining interest keywords are displayed.

3.3 User Vector

The vector for a user is determined from the user's interest keywords and the article vectors as follows.

1. $R_1, R_2, \cdots, R_m$ are article pages read by the user, and these pages have interest keyword j.
2. The vector for the article page R_i is defined by $v_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4})$.
3. μ_{je} is the average value for each element $e(e = 1, 2, 3, 4)$, and σ_{je} is the standard deviation for each element.

$$\mu_{je} = \sum_{i=1}^{m} v_{ie} \Big/ m \,, \quad \sigma_{je} = \sqrt{\sum_{i=1}^{m} (v_{ie} - \mu_{je})^2 \Big/ (m-1)} \tag{2}$$

4. When σ_{je} is less than the threshold, the fluctuations in the element of the vector are small, and the value for that element for interest keyword j is defined as μ_{je}. When σ_{je} is more than the threshold, the fluctuations in the element are large, and the value of that element is defined by a "*don't care*". For example, when σ_{j2} and σ_{j3} are less than the threshold and others are more than the threshold, the user vector of a topic j is determined by (*don't care*, μ_{j2}, μ_{j3}, *don't care*).

4 Selection and Ranking of Articles Based on User Preferences

FNR selects articles from pages gathered using the user's interest keywords, and ranks the selected articles by using co-occurrence keywords and the user vectors.

1. Co-occurrence keyword k is extracted from the pages that have interest keyword j.
2. Value c_{jk} of the co-occurrence is calculated using c_{jk} = {(the number of cooccurrences of j & k) + 1 } / { (frequency of j) + (frequency of k)}
3. Article page P_i, which includes interest keyword j, is selected from m accessed articles by the user.
4. Cosine similarity is computed based on the distance between P_i and c_{jk} based on interest keyword j.
5. If P_i is more than the threshold, P_i is selected
6. The cosine similarity of sentiment D_i is computed based on the distance between article vector $v_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4})$ of P_i and the $v_j = (v_{j1}, v_{j2}, v_{j3}, v_{j4})$ user vector for interest keyword j, which is calculated as described in Section 3.3, and D_i is calculated using

$$D_i = \sum_{e=1}^{4} (v_{ie} \times v_{je}) \Big/ \sqrt{\sum_{e=1}^{4} v_{ie}^2 \times \sum_{e=1}^{4} v_{je}^2}. \tag{3}$$

   However, if v_{je} is more than the threshold, the calculation of v_{je} is excluded because v_{je} is a "*don't care*" term".
7. P_i is displayed if D_i is larger than a threshold.

5 Evaluation

We have developed a prototype FNR.It was developed using Microsoft Visual Studio .Net C# and Perl. The morphological analysis was done using Mecab[8]. This section presents the experimental results obtained with the prototype system and discusses the user vector changes based on the user's browsing history. The articles were collected on April 28, 2005, between 9:00 and 9:30 a.m. from six news web sites: there were a total of 255 articles with metadata. The MPV site categorized and integrated the metadata based on user preferences. The extraction threshold for keywords of interest was set at 0.06 because at least one interest keyword must be extracted from each article; the extraction threshold for words such as proper nouns and general nouns in each article was set at 0.1 because at least 11 words must be extracted from each article.

5.1 User Vector

Figure 2 shows the changes in the average and the standard deviations for our first experiment. Both graphs show the values for each vector element, *e2* to *e4*. In this case, the user selected an article about "topic of *a country*" which has the opinion of agreement or opposition. The user initially selected the article at random and then gradually selected articles expressing opposing opinions. The standard deviations for *e2*, "grad ⇔ sad", and *e3*, "relaxation ⇔ strain" were initially relatively high, as shown in (Fig.2 (a)). Both gradually decreased as the user selected articles. The average value of *e2* was smaller than 0.5(Fig.2 (b)), indicating that the user had selected an article that should create an sad (unhappy) sentiment and that the sentiment was modeled correctly. The other average values of *e1*, *e3* and *e4* showed that the user also selected articles that created "rejection", "strain", and "anger" sentiments.

These results show that FNR can model the sentiments for a user about topics based on the article vectors by using his or her browsing behavior.

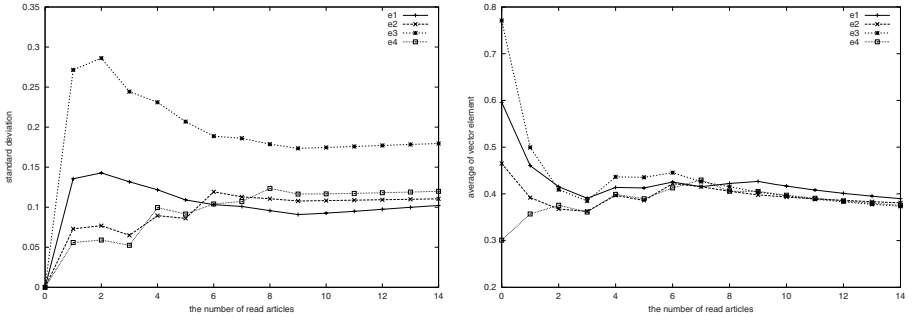

Fig. 2. Changes in average and standard deviation of impression vector based on user's browsing history (keyword of interest is a country name): Changes in standard deviation value (left), Changes in average value (right)

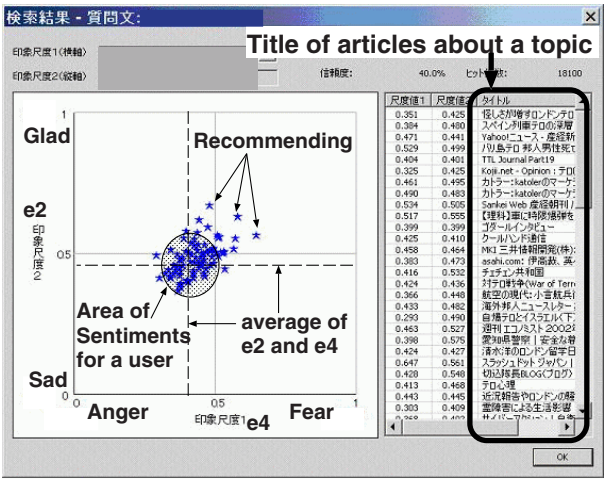

Fig. 3. Plots of sentiments for user and article vector for a topic

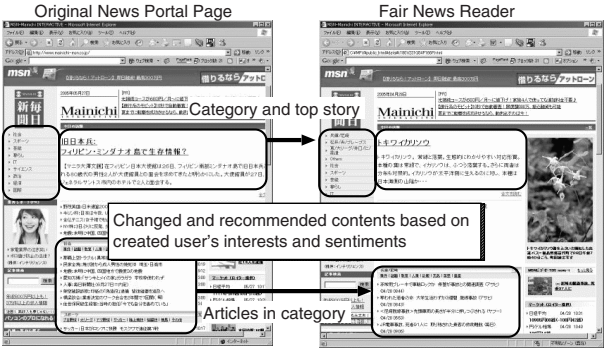

Fig. 4. FNR page showing results obtained using different original news portal page from that in Fig. 2

5.2 Recommendation Based on Sentiments for User

Figure 3 shows a plot of the article vectors and an area of sentiments for a user about a topic of "*a country*" the same as in Fig.2. The dotted lines are the average of *e2* and *e4* of the user's sentiment vector, and the stars of dots represent the articles collected about a topic.

The area of sentiments for the user is represented by circle of center *e2* and *e4*. When the value of the standard deviation was under 0.2, *e2* was 0.47 and *e4* was 0.39. The user had read articles that created "sad" and "anger" sentiments, and FNR recommended articles that created opposite sentiments, such as "glad" and "fear", for balanced reading.

5.3 FNR Prototype

Figure 4 shows an example original news portal page and the FNR page. FNR changed the content in three areas based on the user's browsing of several news pages. First, the original category keyword area was mapped onto the user's interest keywords. Next, the top news article with an image was selected based on the interest keywords and was displayed as a title that has not been read yet. Then, the titles of the article in each category were replaced with those of articles containing the interest keywords and having higher valued vector.

6 Related Work

There has been considerable investigation of portal site technology for gathering, categorizing, integrating, and recommending information.

Columbia's Newsblaster [9] is an online news summarization system in which collected news articles are categorized by event using a topic detection and tracking (TDT) method and TF · IDF. After each news article has been assigned to one of the six categories, each category is summarized using language processing technology. The user can then read a brief summary of an event based on information collected several Web pages. However, the method of article's summarization can not consider about the various sentiments or aspects.

MSN Newsbot[3] uses not only collection and classification technology but also personalization technology. A user's preferred articles are selected using personalized information based on his browsing history. However, the method of article selection is not good enough because the system only adapts to user interest, and the selected articles are not categorized.

Methods of extracting information about writers from movie reviews, book reviews, and production evaluation questionnaires also have been studied. Turney [10] proposed a method of classifying various genres of reviews into "recommended" or "not recommended". His method extracts specific patterns of phrases from input text, calculates mutual information, and takes the difference, where the two reference words were heuristically determined by him. However, using this method it is difficult to satisfy multiple impressions because the two reference words were designed only for a specific impression scale; "recommended – not recommended". We hit on the idea of classifying input documents into two or more impression classifications using a text classification method.

Many researchers have previously tackled the problem of creating more accurate classifiers using less accurate answer data [11,12]. However, they were not successful because their methods required a large amount of correct answer data. Our method, by contrast, can classify documents using only a little correct data, which can be reused as correct answer data, and which may contribute to the creation of classifiers that are sufficiently accurate.

7 Conclusion

We have developed a news portal system called Fair News Reader (FNR) which unbiasly recommends news articles with different sentiments for a user. The

algorithm dynamically determines the sentiments of news articles and the sentimental preferences for a user based on previously read articles, and discovering symmetric articles against the read articles.

We will evaluate FNR more fully by monitoring the browsing behavior of many users, and will adapt the user preferences' modeling to other types of Web sites, such as shop sites.

References

1. NewsCrawler, http://www.newzcrawler.com/
2. Kawai, Y., Kanjo, D., Tanaka, K.: My Portal Viewer: Integration System based on User Preferences for News Web Sites. In: Andersen, K.V., Debenham, J., Wagner, R. (eds.) DEXA 2005. LNCS, vol. 3588, pp. 156–165. Springer, Heidelberg (2005)
3. Newsbot, http://uk.newsbot.msn.com
4. GoogleNews, http://news.google.co.jp
5. Cai, D., Yu, S., Wen, J.-R., Ma, W.-Y.: Extracting Content Structure for Web Pages Based on Visual Representation. In: Zhou, X., Zhang, Y., Orlowska, M.E. (eds.) APWeb 2003. LNCS, vol. 2642, pp. 406–417. Springer, Heidelberg (2003)
6. Nihon Keizai Shimbun, Inc.: Nikkei Newspaper Full Text Database DVD-ROM, 1990 to 1995 editions, 1996 to 2000 editions, 2001 edition, Nihon Keizai Shimbun, Inc.
7. Kumamoto, T., Tanaka, K.: Proposal of Impression Mining from News Articles. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3681, pp. 901–910. Springer, Heidelberg (2005)
8. MeCab (2004), http://chasen.org/~taku/software/mecab/
9. McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J., Sable, C., Schiffman, B., Sigelman, S.: Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster (2002)
10. Peter, D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proc. of Conference on Association for Computational Linguistics (2002)
11. Nagata, M., Taira, H.: Text classification — Trade fair of learning theories. In: IPSJ Magazine, vol. 42 (2001)
12. Tsukamoto, K., Sassano, M.: Text categorization using active learning with AdaBoost. In: IPSJ SIG Notes, NL126-13 (2001)