RELEVANT$^{\text{News}}$: a semantic news feed aggregator*

Sonia Bergamaschi[1], Francesco Guerra[2], Mirko Orsini[1], Claudio Sartori[3], and Maurizio Vincini[1]

[1] DII-Università di Modena e Reggio Emilia
via Vignolese 905, 41100 Modena
firstname.lastname@unimore.it
[2] DEA-Università di Modena e Reggio Emilia
v.le Berengario 51, 41100 Modena
firstname.lastname@unimore.it
[3] DEIS - Università di Bologna
v.le Risorgimento 2, 40136 Bologna
claudio.sartori@unibo.it

Abstract. In this paper we present *RELEVANT*$^{\text{News}}$, a web feed reader that automatically groups news related to the same topic published in different newspapers in different days. The tool is based on *RELEVANT*, a previously developed tool, which computes the "relevant values", i.e. a subset of the values of a string attribute. Clustering the titles of the news feeds selected by the user, it is possible identify sets of related news on the basis of syntactic and lexical similarity.
RELEVANT$^{\text{News}}$ may be used in its default configuration or in a personalized way: the user may tune some parameters in order to improve the grouping results. We tested the tool with more than 700 news published in 30 newspapers in four days and some preliminary results are discussed.

1 Introduction

Many newspapers publish their news in Internet. A recent research from the Italian Institute of statistics[4] shows that there is an increasing trend of mastheads publishing their contents on the Net often joining to the paper edition an Internet edition with special and more complete information [5]. Internet newspapers may update their contents frequently: thus there is not a daily issue but the news are continuously updated and published. As a consequence, hundreds of thousand of partially overlapping news are daily published.

The amount of information daily published is so wide that is unimaginable for a user. On the other hand, the availability of news generates new updated information needs for people. The RSS technology supports Internet users in staying updated: news

* This work was partially supported by MIUR co-funded project NeP4B (http://www.dbgroup.unimo.it/nep4b) and by the IST FP6 STREP project 2006 STASIS (http://www.dbgroup.unimo.it/stasis).

[4] http://www.istat.it

[5] Istat report about the Italian online newspapers (years 2005-2006), available at http://culturaincifre.istat.it/

are published in the form of RSS feeds that are periodically downloaded by specific applications called feed readers. In order to improve the users' selection of the interesting feeds from different newspapers, publishers group feeds in categories.

The RSS technology and the news classification in categories does not solve all the "news overload" issues. First, the categories are not fixed, and then the same topic may be called in different sites in different ways. Consequently, a user that wants to be updated about a specific topic has to manually browse the categories of potentially all the newspapers looking for interesting news. Then, the amount of news feeds daily published is so wide that automatic tools are required. If we consider the feeds published only by the five main Italian newspapers in one day, more than one thousand of news are available in their websites [6]. Such news are partially overlapping, since different newspapers publish the same information in different news. RSS feeds from different newspapers may carry the same information in different places, and therefore can confuse the reader. A great improvement might be to show groups, and leave to the reader the optional task of drilling down the group, if necessary, to compare the different flavours of the same information given by the different sources.

This work relies on *RELEVANT*[7] [1], a tool for calculating the "relevant values" among the string values of an attribute. The tool has been conceived for improving the user's knowledge of the attributes of database tables: by means of clustering techniques, *RELEVANT* provides to the user a synthetic representation of the values of the attribute. In particular, *RELEVANT* takes into account syntactic, dominance and lexical relationships for defining similarity measures among the attribute values. Such measures are then exploited for producing clusters of values, which are related. *RELEVANT* is independent of the attribute domain: a set of parameters allows the user to tune the relevance of the similarity measures and the clustering thresholds in order to produce best results.

In this paper we propose *RELEVANT*$^{\text{News}}$, a web feed reader with advanced features, since it couples the capabilities of *RELEVANT* and of a feed reader. By applying *RELEVANT* to the titles of the feeds, we can group related news published by different newspapers in different times in semantically related clusters. In particular, each cluster contains news related under the following dimensions: 1) Spatial perspective: the news with the similar titles published in different newspapers; 2) Temporal perspective: the news with the similar titles published in different times.

Several feed readers have been proposed in the literature (see section 5 for related works), but at the best of our knowledge *RELEVANT*$^{\text{News}}$ is the only lexical knowledge based feed aggregator.

The outline of the paper is the following: in order to ease the comprehension of our news aggregation technique, section 2 recalls the description of the *RELEVANT* prototype together with a detailed description of our technique for computing relevant values. Section 3 shows the *RELEVANT*$^{\text{News}}$ architecture, and in section 4 we discuss some preliminary results. Section 5 shows some related works and, finally, section 6 introduces future work.

[6] We considered the feeds on average available in the newspapers "Il Corriere della Sera", "La Repubblica", "La Gazzetta dello Sport", "Il sole 24 ore", "La Stampa" in a week of analysis.

[7] See http://www.dbgroup.unimo.it/relevant for more references.

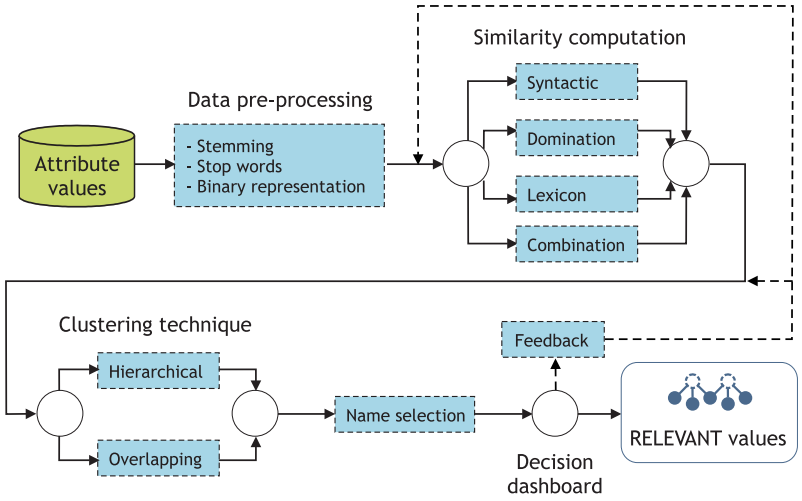

Fig. 1: The *RELEVANT* functional architecture

2 The *RELEVANT* prototype

RELEVANT is based on the idea that analyzing an attribute domain, we may find values which may be clustered because *strongly related*. Providing a name to these clusters, we may refer to a relevant value name which encompasses a set of values. More formally, given a class C and one of its attributes At, a **relevant value** for it, rv^{At} is a pair $rv^{At} = \langle rvn^{At}, values^{At} \rangle$. rvn^{At} is the name of the relevant value set, while $values^{At}$ is the set of values referring to it.

Figure 1 shows the functional flow diagram, including the following tasks:

1. **Data pre-processing**: like most cluster tasks with non-numeric attributes, the problem is to find an effective representation of the points (i.e. the attribute values) in a space, and to devise a suitable similarity function to be exploited by the clustering algorithm. After usual stemming, we build a binary representation of the attribute values, exploiting three different kinds of measures: 1) *syntactic*, mapping all the words of the attribute values in an abstract space, and defining a syntactic similarity function in such space; 2) *dominance*, expressed by the root elements described later on; and 3) *lexical*, which identifies semantically related values expressed with a different terminology.
   The syntactic similarity is based on the assumption that words related to the same object may have the same etymology and then share a common root. Thus, we group different attribute values sharing common words. However, syntactically similar values may refer to different objects. As a consequence, a similarity computation based only on a syntactical method may generate clusters containing elements with different meanings, but in conjunction with the similarities described below, it provides satisfactory results.

A similarity measure may be extracted from the Dominance relationships between the attribute values. Considering two attribute values a_1 and a_2, we say that a_1 dominates a_2 if the meaning of a_1 is more "general" than a_2. Any partial order on attribute values could be used to define dominance. We observed that it is frequent to have string domains with values composed of many words and including abbreviations and that the same word, or group of words, may be further qualified (i.e. specialized) with multiple words in many ways. Thus, we approximate the dominance between attribute values, a semantic property, with the *Contains* function, a syntactic property. *Contains* is a function based on string containment: $Contains(X,Y) = true$ iff $stem(X) \supseteq stem(Y)$, where X and Y are sets of words and *stem* is a *stemming operator*. The dominance is a partial order and can be represented by an oriented graph. Dominance is useful to build clusters of values around *root elements*. A root element is an attribute value with only outgoing edges in the domination graph, and can be taken as a representative of the cluster composed by the nodes recursively touched by its outgoing edges.

Finally WordNet is exploited for providing lexical similarity. In WordNet, English words are grouped into sets of synonyms (synsets), each one expressing a distinct concept. Synsets are described with a definition (a *gloss*) and are interlinked by means of conceptual/semantic and lexical relations. Since a word may be associated to different synsets due to the polysemy, a user is generally requested to manually select the appropriate synset for each term. On the other hand, by exploiting the WordNet lexical similarity it is possible to group different values which refer to semantically related synsets. Two different values, sharing one or more synsets are potentially similar. We can thus compute similarity on the basis of the shared synsets.

2. **Similarity Computation**: two tasks are enabled: the selection of the metrics for computing the similarity between pairs of attribute values and the selection of the similarity measures to be used (syntactic, dominance, lexical or a combination of the three). Concerning the first task, RELEVANT considers the choice of the metrics as a parameter, which can be chosen by the integration designer and changed to compare different settings. The tool implements some of the metrics commonly adopted in information retrieval (Simple Matching, Russel & Rao measure, Tanamoto Coefficient, Sorensen measure, Jaccard's Similarity [10]). Due to the sparseness of the binary matrix, the Jaccard similarity measure, which only considers the positive values in both the attribute value representations[8], is set as default. Concerning the second task, the user may balance the weight of the different similarity measures by setting specific parameters.
3. **Clustering technique**: this module implements some clustering algorithms to compute the set of relevant values on the basis of the selected similarity computation. The designer may choose between a classical clustering algorithm (generating partitions), and an overlapping clustering algorithm to compute *values*. At present we implemented a hierarchical clustering algorithm (see [4]) and an overlapping one

[8] Let us define B_{11} as the total number of times a bit is ON in both bit strings, B_{00} as the total number of times a bit is OFF in both bit strings, and L as the length of the bit string, the Jaccard Measure is defined as $B_{11}/(L - B_{00})$

based on “poles”, which are a rough partition of the domain (see [2]). Poles can be either the output of the hierarchical clustering algorithm or the set of root elements.

4. **Name selection**: The simplest way to detect a list of rvn_i candidates, i.e. the maximal values among *values*, is to use the *Contains* function. The integration designer may select the most appropriate name among them.
5. **Decision dashboard**: the integration designer may interact with *RELEVANT* in two ways: with the simple or advanced mode. The simple mode uses some default parameters and provides four sliders: the first to select the precision of the relevant values set. The slider ranges from *rough*, producing a small number of large relevant values where an attribute value may belong to different relevant values, to *accurate*, generating a large number of small relevant values each of them containing closely related attribute values. The second/third and fourth slider allow the selection of the weight of the similarity computation method. In the advanced mode, the designer may set among about hundred different configurations and clustering thresholds.
6. **Feedback**: The system provides a feedback on the results of a run that may be exploited to refine the relevant values set. A set of standard quality measures allows the tuning activity:
   - **countRV**: number of relevant values obtained for the configuration;
   - **average, max_elements, variance**: the descriptive statistics over the number of elements;
   - **count single**: number of relevant values with a single element;
   - **Rand Statistic index, Jaccard index, Folkes and Mallows index [6]**: compute the closeness of two sets of clusters evaluating couples of values that belong to the same cluster in both the sets;
   - **silhouette [9]** (only if the hierarchical clustering algorithm is used): calculates a width for each cluster based on the comparison of its tightness and separation;
   - **overlapping degree** (only if the overlapping clustering algorithm is used): number of elements which are in more than one relevant value.

   In the simple mode, the tuning activity is automatic: *RELEVANT* iteratively computes sets of relevant values to obtain a set of relevant values according to the slider indication. In the advanced mode, the designer may autonomously change the settings obtaining a set of relevant values satisfying his requirements.

3 *RELEVANT*$^{\text{News}}$ architecture

RELEVANT$^{\text{News}}$ is a web application including three components:

- A **feed aggregator** is in charge of collecting the feeds selected by the user;
- A **RSS repository**: *RELEVANT*$^{\text{News}}$ requires a database for sharing feeds published in different days by different newspapers;
- ***RELEVANT*** computes and groups similar news.

The *RELEVANT*$^{\text{News}}$ functional architecture is composed of four steps (see figure 2):

1. **selection of the news feeds**: a simple graphical user interface allows the user to select the interesting news feeds (by means of their URL) and to setup the updating policy, i.e. how frequently the feed has to be checked for new items;

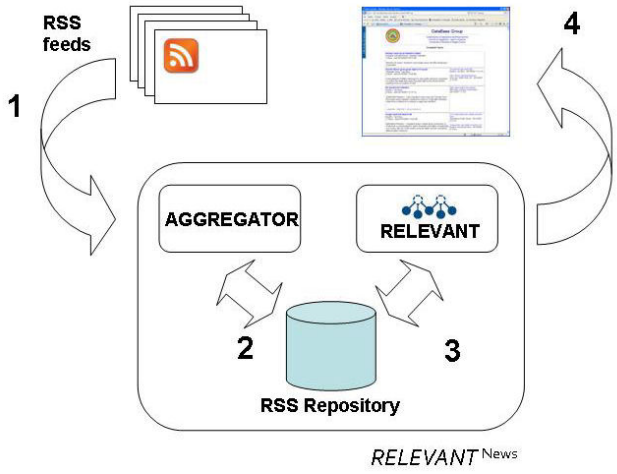

Fig. 2: The *RELEVANT*$^{\text{News}}$ functional architecture

2. **repository population**: a database supports the collection of the feeds. Thus it is possible to provide to the user news that are related to a topic, but are no longer published. The user may select a deadline for the maintenance of the news;
3. **news clustering**: by means of *RELEVANT* similar news are grouped, and for each cluster, a news, representative of the cluster, is selected. Concerning the clustering process, a simple graphical interface may allow the user to parametrize the algorithm settings, establishing the dimension of the clusters (big clusters with loosely related news, or small clusters containing strictly related information), and tuning the weight of the different similarities (lexical, dominance and syntax). Concerning the selection of the news representative of the cluster (the *relevant news*), the user may choose: a) the name extracted by *RELEVANT* ; or b) the last published news;
4. **Relevant news publication**: a web interface shows the news in terms of title, source, date and content. In case of clustered news, the relevant news is visualized together to the list of cluster related news.

In figure 3, a screen-shot of the *RELEVANT*$^{\text{News}}$ interface is shown. Each box contains a different news. In case of similarities, the relevant news text is shown in the box and the cluster related ones can be reached through a link in the bottom box.

4 Preliminary results

We tested *RELEVANT*$^{\text{News}}$ analyzing 730 news from 30 feed providers, published from the 1st to 4th of October 2007. The limited number of feeds allows us to evaluate the results by means of quality indexes provided by the tool, and of some qualitative, user-supplied evaluations. Since a gold standard for news does not exist, and different clusters for the same set of feeds may be provided by a domain expert due to the different

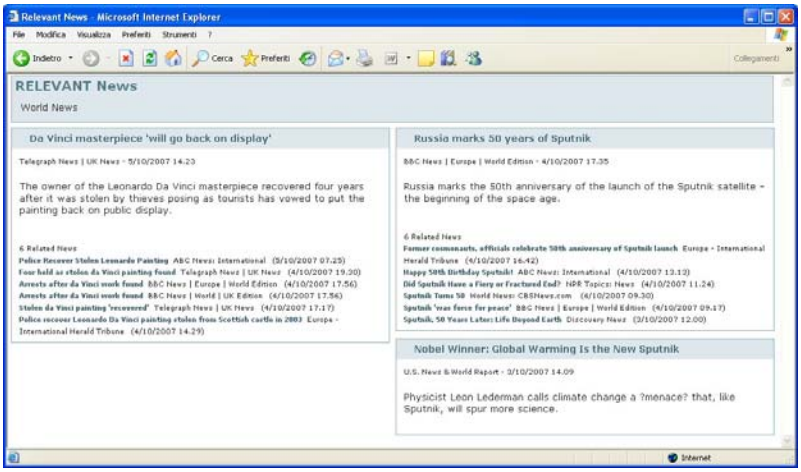

Fig. 3: *RELEVANT*$^{\text{News}}$ screen-shot

grouping criteria may be adopted, in the following, we will analyze several settings producing different clusters of news. Later on, after a brief explanation of the dataset, we will discuss some results and some numerical evaluations of 12 configurations.

4.1 Dataset analysis

The case study is defined by choosing 16 different news publishers with similar RSS feeds topic, i.e. world news, U.S. news and Europe news, analyzing altogether 30 RSS feeds. In particular, we considered 9 newspapers (6 from U.S., Chicago Tribune, New York Times, Wall Street Journal, Time, USA Today and U.S. News, and 3 European, Daily Telegraph, The Guardian and International Herald Tribune), 5 TV Network (4 from U.S., ABC News, CNN, CBS and Discovery Channel, and BBC from U.K.) and 2 International Press Agencies (Reuters and NPR).

During the period of time in analysis (4 days), each publisher provided on average 45 news, with a peak of Daily Telegraph, 141 news and ABC, 89 news, while Chicago Tribune only 5 news. Since the news topics are partially overlapping, the news are also partially overlapping: the same information may be published in different feeds at the same moment.

4.2 Evaluation of the results

Since it is not possible to compare the results produced by *RELEVANT*$^{\text{News}}$ with a gold standard, we will discuss and compare the results computed with different settings. In particular, we considered three different thresholds for the clustering algorithm (since the optimal number of clusters is not known , we considered thresholds producing respectively 300, 450 and 600 clusters) and two different tunings of the similarity parameters. Concerning these parameters, we evaluated a “lexical configuration”, where the

lexical similarity among the news titles assumes a main role, and a "syntactic configuration", where the syntactic similarity is the main similarity measure. We did not take into account a "dominance configuration" as its application in the analyzed dataset is not significant, i.e. only few title are "contained" in other titles.

	SYN_300	LEX_300	SYN_450	LEX_450	SYN_600	LEX_600
# single	241	217	388	348	527	511
Max elem	268	78	30	20	9	7
Avg elem	2.49	2.40	1.55	1.61	1.22	1.24
Variance	16.17	7.31	2.13	1.96	0.81	0.81
Silhouette	0.31	0.34	0.35	0.35	0.45	0.49

Table 1: Qualitative results

The results computed by the *RELEVANT* feedback module are summarized in table 1. The Silhouette values highlight a good clustering process in all the settings (ranges from -1, worst to +1, best). Another interesting information is provided by the "count single" value, that in all the settings is closed to the number of obtained cluster (almost 80% of the computed clusters contains only one element in all the settings). These values, which may be symptom of weakness of the tool are due to news for which no significant similarity has been found. The analysis of the dataset confirms that the observed news are related to general/generic topics from the world, and in the period of time in observation no event with a worldwide importance happened. Thus, we may suppose that clusters similar to the ones computed by $RELEVANT^{News}$ may be produced by a human reader.

In table 2, the clusters obtained considering two different configurations are analyzed. Qualitative analysis shows that lexical similarities improve the results. Table 2.a, where the clusters are computed with the syntactic configuration, shows that the news related to the recover of a stolen Leonardo da Vinci painting are grouped in two clusters. On the other hand, in Table 2.b the news are grouped in the same cluster, due to the lexical similarities among the news titles. Similar considerations may be done for news titles represented in Tables 2.c and 2.d. In this case, it is interesting to observe that the syntactic configuration produces three clusters, but the first news is correctly not included in a overall cluster (see Table 2.d) in the lexical configuration, since it refer to a different topic.

5 Related Work

There is a rich literature about metadata extraction and clustering techniques both in the area of Semantic Web, where metadata support automatic applications to understand web-site contents and in the area of Information Retrieval, where they allow document classification. For some references about this topic, see [1].

Several aggregators have been developed and implemented. Most of them are available as commercial products and their internal mechanisms are not known [9]. It is possible to group them in three different categories:

[9] See http://www.dmoz.org/Computers/Software/Internet/Clients/WWW/Feed_Readers/ for a non complete set of aggregators.

(a) News related to the "Da Vinci" stolen grouped with the syntactic configuration

#1	Arrests after da Vinci work found
	Da Vinci masterpiece "will go back on display"
	Four held as stolen da Vinci painting found
	Stolen da Vinci painting "recovered"
#2	Police recover Leonardo painting stolen from Scottish castle in 2003
	Police Recover Stolen Leonardo Painting

(b) News related to the "Da Vinci" stolen grouped with the lexical configuration

#1	Arrests after da Vinci work found
	Da Vinci masterpiece "will go back on display"
	Four held as stolen da Vinci painting found
	Stolen da Vinci painting "recovered"
	Police recover Leonardo painting stolen from Scottish castle in 2003
	Police Recover Stolen Leonardo Painting

(c) News related to the "Sputnik" grouped with the syntactic configuration

#1	Nobel Winner: Global Warming Is the New Sputnik
#2	Did Sputnik Have a Fiery or Fractured End?
	Former cosmonauts, officials celebrate 50th anniversary of Sputnik launch
	Happy 50th Birthday Sputnik!
	Sputnik "was force for peace"
#3	Russia marks 50 years of Sputnik
	Sputnik Turns 50
	Sputnik, 50 Years Later: Life Beyond Earth

(d) News related to the "Sputnik" grouped with the lexical configuration

#1	Nobel Winner: Global Warming Is the New Sputnik
	Did Sputnik Have a Fiery or Fractured End?
	Former cosmonauts, officials celebrate 50th anniversary of Sputnik launch
	Happy 50th Birthday Sputnik!
#2	Sputnik "was force for peace"
	Russia marks 50 years of Sputnik
	Sputnik Turns 50
	Sputnik, 50 Years Later: Life Beyond Earth

Table 2: A clustering example

1. **Simple readers** provide only a graphical interface for visualizing and collecting RSS feeds from different newspapers. Simple functions supporting the user in reading are provided (e.g. search engine, different ordering, association of news to a map, ...);
2. **News classifiers** show the news classified on the basis of criteria sometimes decided by the user. Simple classifications may exploit the categories and/or the keywords provided by the web sites;
3. **Advanced aggregators** provide additional features for supporting the user in reading, clustering, classifying and storing news.

There are several interesting proposals of advanced aggregators in literature. In [5], Velthune, a news search engine is proposed. The tool is based on a naive classifier that classifies the news in few categories. Unlike this approach, *RELEVANT*$^{\text{News}}$ computes clusters of similar news on the basis of their title. Classifying thousands of news in few categories produces large sets of news belonging to the same category that are not easily readable by a user. In [7] the authors propose an aggregator, called RCS (RSS Clusgator System), implementing a technique for temporal updating the contents of the clusters. NewsInEssence [8] is an advanced aggregator that computes similar news on the basis of a TF*IDF clustering algorithm, and provides to the reader a synthesis of them. Although *RELEVANT*$^{\text{News}}$ does not provide any synthesis, it implements a parametrized clustering algorithm based on syntactic/lexical/dominance relationships, that may be properly tuned for improving the creation of the clusters. Finally, the idea of

RELEVANT$^{\text{News}}$ may be compared with Google News [10] where each news is associated with a list of related information. Differently from us, Google News does not allow the user to select the newspapers. All the newspapers are analyzed and the news are provided to the user on the basis of a collaborative filtering [3].

6 Conclusion and future work

In this paper we proposed *RELEVANT*$^{\text{News}}$ a news feed reader able to group similar news by means of data mining and clustering techniques applied to the feed titles. As usual in data analysis, the startup phase requires the setting of several critical parameters. Nevertheless, for a given parameter setting, the technique calculates the relevant news without any human intervention. Moreover the parameters and similarity metrics selection determine the quality of the relevant value news. Therefore, the designer has to carefully evaluate the results and possibly change some parameters in order to improve the result quality.

Future work will be addressed on developing new techniques suitable for the feed domain. In particular, the preliminary results demonstrated that the dominance is a too restrictive condition: it is not frequent for news titles to be contained in other news titles. Moreover, some other techniques to compute the similarity may be exploited. For example, we are studying a similarity based on term frequency-inverse document frequency (TD*IDF), which takes also into account the word-spread. The idea is that unusual words and specific terms may be related to the same news.

References

1. S. Bergamaschi, F. Guerra, M. Orsini, and C. Sartori. Extracting relevant attribute values for improved search. *IEEE Internet Computing*, pages 26–35, Sep-Oct 2007.
2. G. Cleuziou, L. Martin, and C. Vrain. PoBOC: An overlapping clustering algorithm, application to rule-based classification and textual data. In *Proceedings of the 16th ECAI conference*, pages 440–444, 2004.
3. A. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In Williamson et al. [11], pages 271–280.
4. B. S. Everitt. *Cluster Analysis.* Edward Arnold and Halsted Press, 1993.
5. A. Gulli. The anatomy of a news search engine. In Allan Ellis and Tatsuya Hagino, editors, *WWW (Special interest tracks and posters)*, pages 880–881. ACM, 2005.
6. M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17(2-3):107–145, 2001.
7. X. Li, J. Yan, Z. Deng, L. Ji, W. Fan, B. Zhang, and Z. Chen. A novel clustering-based rss aggregator. In Williamson et al. [11], pages 1309–1310.
8. D. R. Radev, J. Otterbacher, A. Winkel, and S. Blair-Goldensohn. Newsinessence: summarizing online news topics. *Commun. ACM*, 48(10):95–98, 2005.
9. P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65, 1987.
10. C. J. van Rijsbergen. *Information Retrieval.* Butterworth, 1979.
11. C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, editors. *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007.* ACM, 2007.

[10] http://news.google.com/