Information Integration in News Articles from Various Sources

Michal Holub*
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
holub@fiit.stuba.sk

Abstract

Today almost everything can be found on the Web. The problem is that information is dispersed across many different sources. Search engines help in finding required information. However, they fail to present relationships. The next step therefore is to automatically integrate relevant information from various sources and present them to the user in a unified manner. In this paper we present our contribution to integration of news articles into groups referring about the same fact. We integrate from various web portals. This can be used for creation of personalized news portal which will keep the user up-to-date about current affairs.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing); H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

information integration, clustering, keywords extraction, text processing, news articles, events

1. Introduction

The Web is a huge information base containing a large portion of human knowledge. This information can be found in various documents which are usually published as web pages. People often rely on this base when searching for the answers to their information needs or questions.

One difficulty with finding the right information is that there is no web page containing all of the information, not event about one particular theme or topic. Users therefore need to use search engines to locate the right documents which may contain the information they are looking for. However, the answer to a question a user has can be composed of many small pieces of information which are scattered across a lot of web pages.

Search engines are effective in locating the documents which contain the keywords from the query. It is then up to the user to manually look through these documents, find relevant pieces of information and join them together in order to get a broader view of the problem. Search engines fail in finding relationships among information entities contained in the documents. They are unable to integrate the information and present a compilation of results to the user; they only sort the documents according to their relevancy against the query. This issue can be addressed with automatic information integration over various sources [9].

Faceted browsing [11] can help to solve this issue of getting an overview of a certain domain. However, the user still needs to formulate the initial query. There are times and domains where this is difficult, or event impossible, to accomplish. Take for example actual world events. The user cannot be aware that there is something happening in the other part of the world right now, and that he should search for more information about it. News portals add hundreds of new articles every day. It is therefore difficult to stay informed about important issues. Integration of information from various news portals and presenting them to the user as a compilation of main events can help solve this problem [10].

In this paper we present our contribution to integration of news articles according to the event they inform about. We group articles from various sources, i.e. various news portals. Our goal is to group articles which are written on the basis of the same event happening at a precise time of the day (e.g. a particular concert or a political meeting), not a broader topic which can be actual for more days or weeks (e.g. music or some politician in general). This task is often done manually by various subjects as news monitoring service.

We present attributes of articles based on which this integration can be done automatically. This can be used to create a news portal automatically aggregating and presenting articles from various websites. It would give the user greater overview of the current affairs and the way

*Doctoral degree study programme in field Software Engineering. Supervisor: Professor Mária Bieliková, Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, STU in Bratislava. Work described in this paper was presented at the 7th Student Research Conference in Informatics and Information Technologies IIT.SRC 2011.

Holub, M. Information Integration in News Articles from Various Sources. Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on Student Research in Informatics and Information Technologies, Vol. 3, No. 2 (2011) 42-46

various journalists refer about it. We also discuss possible extension of the principles applied in the news articles integration to the general information integration on the Web.

The rest of this paper is organized as it follows. In section 2 we present relevant work done in the area of information integration with the focus on news articles. In section 3 we propose a method for news articles integration. Section 4 contains experiments we conducted and their results. Finally, section 5 concludes the paper with discussion and proposals for future work.

2. Related Work

More and more popular form of integrating data from different sources on the Web is creating mashups. A mashup is a combination of things that were initially not meant to cooperate in order to achieve a new goal [1]. A common example of a mashup is to display additional information on the map background (e.g. the location of restaurants or bus stops). Many web portals and services provide APIs which allow users to incorporate their functionality to new applications. Using these APIs we can put more services together to create a new service (a mashup). Users do mashups mainly for aggregating data from more sources, creating alternative user interface for a web application, personalizing certain website or monitoring of certain data source in order to detect changes in it [12].

A use case scenario involving integration of information on the Web is proposed in [2]. The authors describe a blog containing movie reviews. In this scenario each review is automatically enhanced with information from other sources like pictures from movie's official website, biographies of the actors and director from an encyclopaedia or schedules of local cinemas playing the movie. For this to accomplish we need well annotated data sources which provide computer access to their data. One of the biggest challenges is to persuade the owners of websites to include more semantics in them by wider adoption of standards like RSS and RDF.

Some work focusing on integration of data in the domain domain of news articles has been done recently. In [5] various clustering algorithms are evaluated using different similarity measures on the data set containing thousands of news articles. According to the results the most suitable algorithm for articles clustering is k-means with cosine similarity measure. However, it is not clear whether the articles in one cluster are about one broader topic with each serving for further reading. In this case the articles from one cluster would be suitable for recommendations to the users who read at least one article from the cluster and want to know more on that topic. On the other hand, if the articles in one cluster are more lookalike and they describe the same affair using different words, reading of one of the articles is sufficient.

Recommendation of news based on comparison of their content was done in [8]. Authors introduce a feature vector describing each article. This vector contains for example keywords extracted from the title and from the body of the article, article's readability index or names and places mentioned. Using this vector and cosine similarity measure articles from one news provider were clustered in order to recommend articles for further reading. Articles in one cluster shared a broader topic (or cause). Their representation of article might also be used for our purpose, but the grouping algorithm would have to compare the selected features differently.

Another approach to grouping of similar articles is to use a tree structure [13]. Articles are represented as leaves of a tree. Similar articles share a common parent node which summarizes their features, mainly keywords. Again, similarity of two articles is considered as a proposal for further reading, articles are related by one broader topic. This approach can be used in real time similarity computation as every newly created article can be inserted to a proper subtree. This action then triggers further computation and possible tree rearrangement. The tree structure is useful also for recommendation of articles from one category.

A method for topic analysis and detection of categories of news articles is presented in [6]. It can work on-line as soon as new article is published. The method is based on extraction of noun phrases and can work without any document corpus. In order to extract noun phrases, morphological analysis identifying word stems and part of speech tagging is done. One article can be assigned to more than one category, the relevancy against each category may differ. The authors tested their method on 800 English and Japanese articles with results of precision and recall over 90 %. Again, this method can be used for recommending articles from the same category, as a way of further reading proposal.

The most similar work to ours is the Europe Media Monitor project [3, 10]. It is a European Union project in which about 100,000 articles from about 2,000 sources are processed every day. The articles are written in approximately 50 different languages. Gathering of the articles is done via RSS as well as via automatic text extraction from HTML documents. Data is then used by various news and trends monitoring services like NewsBrief[1].

NewsBrief is a service which focuses on detecting breaking news and short-term trends. It clusters articles based on the event which they refer to. One cluster shares a common title and lists hyperlinks to articles from various sources related to the event described in the title. The title for the whole cluster is taken from the most representative article, which is the cluster's medoid. Every 10 minutes this service retrieves the latest articles from 4 hour window and clusters them. At first, every article is considered to be a cluster. Then the most similar clusters are iteratively joined together until certain threshold is reached. For representing the article the word count vector is used together with cosine similarity as the similarity measure. However, sometimes this service puts unrelated articles in the same group. Therefore we see a possibility for improvement in completing this task.

3. Grouping of News Articles

Clustering can be considered as a form of information integration. It is also a common approach when grouping news articles. It takes various entities of the same type and tries to partition them to predefined number of clusters. Clustering algorithms use various vector representations of articles. In the first step, randomly chosen items are set as centres of the new clusters. Then each item is

[1] http://www.newsbrief.eu/

put to the most appropriate cluster based on some similarity metrics. There are many similarity metrics, e.g. cosine similarity, Jaccard index, Pearson correlation coefficient, etc. The next step in the clustering algorithm is to compute new centres of clusters and rearrange the items in clusters. This process repeats until the point in which the change between two consecutive iterations is below a certain limit.

Using clustering is not usable in order to group articles in a way we described. One problem is that we need to have all the articles available at once and we need to know the number of clusters we want to produce (or we can gradually try different numbers and evaluate each result produced). This is not the case of news articles, which are published in different times from different sources. In the ideal case we acquire the article as soon as it is published and decide, whether we already have an article (or a group of articles) referring to the same affair or event. If so, we should add the article to the existing group, otherwise a new group should be created. In this situation we do not know the number of groups we need.

Another particularity in our scenario is that we consider only a small fraction of articles for comparison. Usually, an event happens some time during the day and news articles are published within a day or two from the event. Our feature space is therefore not very large and we can consider more elements from the article in the computation.

3.1 Grouping Algorithm

We group articles from various Slovak news portals in order to get groups of the articles referring about the same affair published within one day. It is a kind of press overview. Each of the portals provides the RSS feature for accessing newly published articles. We subscribe to this RSS channel and periodically download newly created articles. As soon we get a new article we try to put it into a group to other articles which refer to the same event. We distinguish the articles according to their URL, which is unique.

In order to compare the articles we compute the similarity of their notable features. The more features two articles share the more likely they refer about the same event. We experimentally set a threshold for the number of features two articles have to share in order to put them in the same group. Different feature types can also have different weights.

In our method we take keywords extracted from each article as features describing it. For this purpose we use a web service which is a part of the Metall[2] project developed at our faculty [4]. This web service takes a URL of a document or its plain text as the input and returns a set of extracted keywords (which could also be tags or named entities) in English language.

At first, the text from the main content of the given webpage is extracted. This is done to cut off the menu bar and other parts of the web page which do not contain words related to the main content. Then the plain text is translated from the source language to English using Google

[2]http://peweproxy.fiit.stuba.sk/metall/

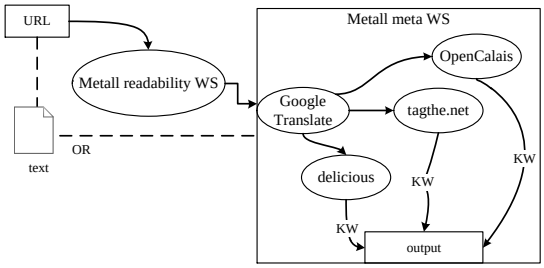

Figure 1: Web service from Metall project for keywords extraction.

Translate[3] web service. This step is necessary because the web services for keywords extraction that we use work only with English language. Afterwards, the translated text serves as an input for various web services like OpenCalais[4], tagthe.net[5] or delicious[6]. These services return keywords or tags which they extracted from the text. The keywords returned are already lemmatized and stemmed so we do not need to do any further processing. The scheme of the Metall web service is depicted in Figure 1.

We do not consider any relationships among extracted keywords, or their relationships to other concepts. The keywords extracted using these services usually contain few named entities (these define what the article is about) and some verbs referring to the named entities (these define the actions connected to named entities). In most cases this is a good representation of an article. One problem can occur when there are different events containing the same entity in one day (e.g. the president is mentioned in two different contexts). In this case the keywords could indicate that the articles are about the same event, although they are not. Possible solution would be to consider also time when the article was published.

When we have the representation of a newly published article we can put it into a suitable group. This is done according to Algorithm 1. We compare the keywords of the newly created article with the keywords of every article which is already in some group and we compute the intersection of these two sets. Then we compute the degree of analogy S of article a against article b according to Equation 1:

$$S = \frac{|kw(a) \cap kw(b)|}{|kw(a)|} \tag{1}$$

where $kw(x)$ is the set of keywords extracted from the article x. We use only the number of keywords of the article a in the denominator because we compute how one article resembles another one, not their mutual similarity (in both directions). This is important because we can have two articles which cover the same event to different degree. Consider that the first article informs about the event in detail, while the second article is brief. The similarity of the second article against the first will be high, but the similarity of the first article against the second will be lower because it contains more information.

[3]http://translate.google.com/
[4]http://www.opencalais.com/
[5]http://tagthe.net/
[6]http://www.delicious.com/

Algorithm 1 Put new article n into suitable group

```
1: article = empty group
2: for all grouped article a do
3:    compute number k of identical keywords of n and
      a
4:    resemblance = k / number of keywords of n
5:    if resemblance > max then
6:       max = resemblance
7:       article = a
8:    end if
9: end for
10: put article n into group of article
```

In the setup phase we give the maximum resemblance a certain threshold value. This is the lowest number of equal keywords in order to consider two articles being about the same affair. If we do not find any suitable group for an article we consider it to be the first article referring about a new event (about which no news has been published yet). We then put this article into a new group. This algorithm is repeated whenever a new article from any monitored source is published. We can see that the number of comparisons increases with increasing number of articles in our dataset. We expect to maintain articles from 1 day window in the production mode.

Each article is put into exactly one group. It is not desirable to have more groups associated with an article as the group defines the resemblance of articles. If there are two suitable groups for one article these should be joined together.

News articles do not always have to inform about an event. They can contain some analysis or opinion of the journalists. These articles would not have counterparts on other news portals as they are unique from their. They form single element groups.

4. Experiments

In the algorithm for grouping of news articles described in the previous section we used the threshold of keywords overlap in order to determine whether two articles refer about the same event. This threshold is the minimal amount of keywords which have to be the same for both articles in order to put them to the same group. In the setup phase of the experiments we manually downloaded and grouped articles from three Slovak news portals. Then we extracted keywords from articles in each group and observed their properties. We found out that articles belonging to the same group had at least 60 % of the keywords in common. On the contrary, articles from any two different groups had at most 20 % of the keywords in common. Based on these empirical findings we expected the ideal threshold value to be 40 %, i.e. two articles have to share at least 40 % of their keywords in order to refer about the same event. We experimented with three different values of the threshold and confirmed that 40 % gave the best results.

As a baseline for our experiments we chose the NewsBrief service which clusters news articles from 4 hour window. It works with articles in many languages, however, we were interested only in slovak news articles. From the NewsBrief website we downloaded clusters of Slovak news containing links to articles. Then we took all links from all clusters, sorted them randomly and used them as an input for our method of grouping articles. We iterated through the whole collection and grouped articles according to algorithm described in previous section. Although we had all articles at once, we processed them as they would be written in different times (first article processed while iterating through the collection was considered to be written sooner than the second article, and so on). This was done to simulate the real conditions in which articles are be published in different times and processed as soon as they are downloaded.

We downloaded 106 articles which were split by NewsBrief into 24 groups. The smallest group contained only 2 articles; the largest group contained 17 articles. We manually evaluated the quality of the grouping, i.e. we checked if the groups contained articles referring about the same event. The largest group was not constructed correctly because it contained articles about 8 completely different events. The largest group which was correct contained 7 articles. There were also some smaller groups which contained an article not related to other articles in that group. Because the baseline used was not 100 % correct we also compared the results of our method with a dataset created by manually correcting errors in the baseline.

We used these 106 articles as an input to our system. The results for 3 different values of threshold are summarized in Table 1. The row with *groups according to baseline* states the number of groups which were constructed equally by our method and by NewsBrief service. The best value of this indicator was achieved when using 30 % threshold.

Articles in wrong groups row states the number of articles which were put into a group to which they evidently did not belong (all groups were also manually evaluated by the human expert). Some groups which we got as a result of our grouping method contained more articles about 2 different events. These groups could be further divided if higher threshold is used. Actually, when using 50 % as a threshold there were no groups which could be further divided. On the other hand, some articles about the same event were dispersed into few groups. This happened because the threshold was set too strict and only very similar articles were put together. The resulting groups can be joined together. Using 40 % threshold was a good compromise between too many and too few groups.

As we already mentioned, the clusters of articles according to the baseline were not perfect. There were articles which were in wrong groups. Our method has put some of these articles into correct groups (the best result achieved with 40 % threshold but the differences are low). Some groups in in the baseline incorrectly contained more articles about various events. Using our method we successfully divided these groups and put the articles into new ones. We achieved the best results of this indicator when using 40 % threshold.

As we can see from the overall results the optimal value of the threshold for the proposed method is close to 40 %. When using the value of 30 % we got the most groups according to the baseline, but we also put many articles into incorrect groups. When using 50 % value of the threshold we got a lot of small groups which could be joined together. This value turned out to be too strict.

Table 1: Results of articles grouping.

	30 % threshold	40 % threshold	50 % threshold
Total groups	34	46	60
Groups according to baseline	19	10	4
Articles in wrong groups	11	4	4
Groups which could be divided	4	1	0
Groups which could be joined	6	18	43
Articles moved to good group	3	4	2
New correct groups made from wrong	4	12	8

5. Conclusion and Future Work

We have presented a contribution to grouping of news articles according to event they inform about. We found out that keywords representing the article are a good property which can be used to do this. Their proper extraction is therefore crucial for the whole process. We can also use term frequencies together with various similarity measures. Additional metadata can refine the results and help deal with ambiguities. Such metadata can be for example time when the article was written. We assume that articles referring to the same event are written in short time one after another. Other metadata can be the category of the article as many news portals put articles into common categories like economics, science or sport.

The approach presented in this article can be generalized to finding relationships among other types of information entities (i.e. web objects). The key is a good representation of an object, which means good attribute and feature detection. Most of the objects are of textual representation; therefore we can use keyword extraction as described in this paper.

In the future work we would like to use this method on a personalized news portal. It will serve as an overview of the actual events. Groups of articles will be presented to the user who can then select the source of the article. This will give the reader an overview of what is happenning right now.

The portal will provide the user with the news as soon as they are published. Today, when the reader reads news only through one favourite portal, he has to wait until the journalists of this portal publish an article about the actual event. Other portal might have published some information sooner and using our solution the reader will be informed faster.

We also plan to incorporate other techniques such as automatic articles categorisation. Thus, the groups of articles will be presented in thematic categories. The news portal will include social personalization and adaptation techniques in order to provide users with news articles suiting their interests. For this purpose we plan to utilize implicit interest indicators in order to evaluate interests of the user, as we described in [7].

Acknowledgements.
This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11, and it is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

References

[1] A. Alba, V. Bhagwan, J. Grace, D. Gruhl, K. Haas, M. Nagarajan, J. Pieper, C. Robson, N. Sahoo. Applications of Voting Theory to Information Mashups. In: *Proceedings of the 2008 IEEE International Conference on Semantic Computing*, pages 10–17. IEEE Computer Society, Washington, DC, 2008.

[2] A. Ankolekar, M. Krötzsch, T. Tran, D. Vrandecic. The Two Cultures: Mashing Up Web 2.0 and the Semantic Web. In: *Proceedings of WWW '07 – 16th International Conference on World Wide Web*, pages 825–834. ACM Press, New York, 2007.

[3] M. Atkinson, E. Van der Goot. Near Real Time Information Mining in Multilingual News. In: *Proceedings of WWW '09 – 18th International Conference on World Wide Web*, pages 1153–1154. ACM Press, New York, 2009.

[4] M. Barla. Towards Social-based User Modeling and Personalization. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, 3(1): 52–60, 2011.

[5] C. Bouras, V. Tsogkas. Assigning Web News to Clusters. In: *Proceedings of ICIW '10 – Fifth International Conference on Internet and Web Applications and Services*, pages 1–6. IEEE Computer Society, Washington, DC, 2010.

[6] D.B. Bracewell, J. Yan, F. Ren, S. Kuroiwa. Category Classification and Topic Discovery of Japanese and English News Articles. *Electronic Notes in Theoretical Computer Science*, volume 225, pages 51–65, 2009.

[7] M. Holub, M. Bieliková. Estimation of User Interest in Visited Web Page. In: *Proceedings of WWW '10 – 19th International Conference on World Wide Web*, pages 1111–1112. ACM Press, New York, 2010.

[8] M. Kompan, M. Bieliková. Content-Based News Recommendation. In W. Aalst, J. Mylopoulos, N.M. Sadeh, M.J. Shaw, C. Szyperski, F. Buccafurri, G. Semeraro, editors, *E-Commerce and Web Technologies*, volume 61 of *Lecture Notes in Business Information Processing*, pages 61–72. Springer, 2010.

[9] M. Raza, F.k. Hussain, E. Chang. A Methodology for Quality-based Mashup of Data Sources. In: *Proceedings of iiWAS '08 – 10th International Conference on Information Integration and Web-based Applications & Services*, pages 528–533. ACM Press, New York, 2008.

[10] R. Steinberger, B. Pouliquen, E. Van der Goot. An Introduction to the Europe Media Monitor Family of Applications. In F. Gey, N. Kando, J. Karlgren, editors, *Information Access in a Multilingual World – Proceedings of the SIGIR 2009 Workshop*, pages 1–8. Boston, USA, 2009.

[11] M. Tvarožek. Exploratory Search in the Adaptive Social Semantic Web. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, 3(1): 42–51, 2011.

[12] J. Wong, J. Hong. What do We "Mashup" When We Make Mashups? In: *Proceedings of WEUSE '08 – 4th International Workshop on End-user Software Engineering*, pages 35–39. ACM Press, 2008.

[13] D. Zeleník, M. Bieliková. News Recommending Based on Text Similarity and User Behaviour. In: *Proceedings of WEBIST '11 – 7th International Conference on Web Information Systems and Technologies*, pages 302–307. Noordwijkerhout, The Netherlands, 2011.