Information Extraction from plain Text in News

Maheshkumar Nanda

Student, Masters in Computer Engineering, K. J. Somaiya College of Engineering, Mumbai

Abstract- With the advancement in technology, various options to get latest news updates have increased significantly. News updates are easily available on web. So many users have shifted from traditional newspaper to digital news. This poses serious challenges to news providers where they need to make latest news readily available to user at all instants. Our project deals with similar issues and recommends news providers what exclusive news they are missing when compared to their competitors.

Index Terms- News, Collection, Similarity, Keywords, Stop words, Ranking.

I. Introduction

1.1 **Overview**

Information retrieval and extraction techniques can be efficiently applied along with latest web technologies available to develop many applications beneficial to various businesses as well as users. In our system we develop two modules. First module assimilates news from various sources to process news. It has sub modules which gather information, find similarity in news, find the differences in news, removes stop words, recommends a particular news agency the news they are missing when compared to other news channels. Second module after collecting news from a source facilitates user to view news of that news channel in a language which user desires from a list of languages. So even News providers of local languages will be able to use this project.

1.2 Motivation

Today being a world where every moment a new News is created, every person irrespective of their language and community would like to be updated with latest happenings in the society. With technological advancement various gadgets have become too common and are easily accessible to people. One would like to read news from a particular source in some other language. But he may not be able to do so as he is not familiar with the language. So we can develop an application which collects news in a language specified by news agency and parse it to convert to language desired by the user. User can choose amongst list of languages which is given. For Example. If a user from village wants to read what news BBC world is publishing on their website. But he may not be able to do so as he doesn't know English Language.

Many applications have been already developed in news domain for the users need. But the news providers also want automated processes to facilitate user easiness. They require to compare their news with their competitors for smooth flow in business and not missing out any exclusive news. Most of these tasks are done by human analysts. We can provide an application which helps news provider to compare their news with competitors. We can also recommend what they are missing out in their news so that they have exclusive coverage of latest news.

1.3 Scope

System proposed could be further remodeled and refined for servicing complex user requirements. Our system would only process Text news. News in images and video will be out of scope of this project. Also we can limit news sources to 5 which can be further extended to 50 news sources if complex system is desired. Our system would only recommend keywords relevant to missing news and not the entire sentences. TextRank algorithm and idf scores will be used to rank keywords at various levels. Other ranking schemes are not considered. Also we have not taken time when news has been published online into consideration as "top stories" from various sources are retrieved which contain news updates related to all domains. Source news channels considered would be in English language as English is a standard language accepted worldwide. It can be extended to translate from any language to any other language. We will recommend missing keywords to news sources. System can be also designed to recommend entire sentence and further also the domain from which they can add up news in their source.

II. Information Extraction from plain Text in News

2.1 News Representation

When talking about News collection, first issue which may arise is what kind of news representation do we require for our system? It can be text, audio, image or any other format. Our system would be limited to news in text format. The news present on web is mostly in XML format. So in our application news would be text retrieved from XML which is embedded in between tags.

2.2 News Collection

News in our system would be collected from various online sources. Various technologies are available for retrieving news from online sources. News can be efficiently gathered from different sources using RSS [4]. Different news websites allow RSS feeds to be included in code so that real time news data can be fed dynamically to the application for processing. Only the data within the title tag of sources can be stripped off to collect the headlines [6] as headlines are of more importance in our application as compared to entire article because we need to spot the differences and similarities in news and not the quality of content.

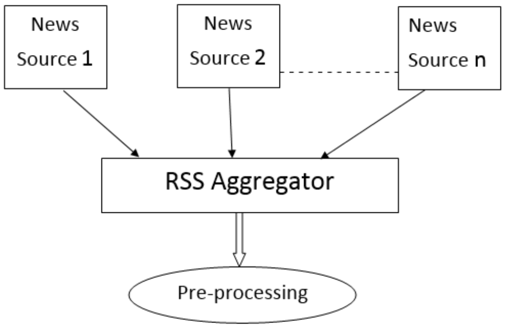


Figure 1. Pre-processing

2.3 Removal of Stop words

Feeds from different sources may contain irrelevant information. But as we are just stripping content of title tag, most of the content is relevant. Still it will contain many stop words like at, a, the, is, above, under etc. which need not be compared otherwise which may lead to wrong calculation of distance measure or similarity measure. Only keywords other than stop words need to be compared in multiple sources. Term weighting can be used for calculating similarity measure. It requires counting of keywords which are same and accordingly news can be weighted. This news is recommended to news agency only if weight is above threshold value. Keywords can also be ranked. And only those keywords above a certain rank need to participate in recommendation [10].

2.4 Similarity Check

Concept which other methods use like tf for weighting the words in a document, cannot be used here because we are only extracting information in title [1, 4].

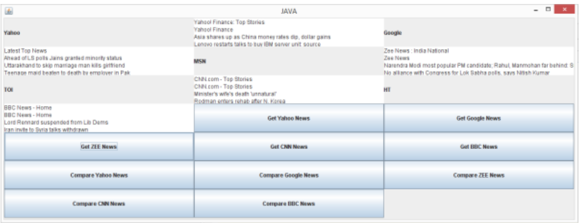
idf = log(number of documents/no of documents containing that term) as given in equation (2).

Even idf {inverse document frequency} may give false measure. If the number of sources in which a particular keyword occurs is more it means that keyword is present in most news sources and rest news sources are missing on that story. But in case of exclusive news inverse of idf would be helpful as only one source has breaking news and rest sources are missing out that news. We neglect the latter case assuming in this competitive world most of the channels would have exclusive story on their page.

III. Implementation

3.1 Get News

The first requirement of our implementation should be acquiring news from various sources. We use a very simple technique of RSS feeds. The headlines from various sources are extracted using RSS and fed to the GUI which news providers can view. Following snapshots make it clear that by clicking buttons for respective news channels, news of corresponding channel can be acquired in textbox.



Yahoo! Finance: Top Stories
Yahoo! Finance
Asia shares up as China money rates dip, dollar gains
Lenovo restarts talks to buy IBM server unit: source

Latest Top News
Ahead of LS polls Jains granted minority status
Uttarakhand to skip marriage man kills girlfriend
Teenage maid beaten to death by employer in Pak

Zee News : India National
Zee News
Narendra Modi most popular PM candidate; Rahul, Manmohan far behind: S
No alliance with Congress for Lok Sabha polls, says Nitish Kumar

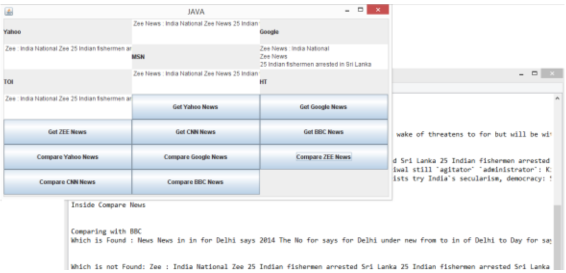
CNN.com - Top Stories
CNN.com - Top Stories
Minister's wife's death 'unnatural'
Rodman enters rehab after N. Korea

BBC News - Home
BBC News - Home
Lord Rennard suspended from Lib Dems
Iran invite to Syria talks withdrawn

Fig 2: Get All News

3.2 Compare News

Also we have compare news button which can be set to show either the similarities in news of various channels or the difference between them. Coding varies accordingly. While showing the similarities or the differences it is very important that less significant words [12] in the news are not compared. Hence we remove stop words which do not participate in similarity check. Not all keywords should be checked. So we can rank [12] the keywords according to idf {inverse document frequency) as discussed before. Other methods of ranking keywords like position based ranking, graph based ranking and centroid based technique can also be used [7]. Only those keywords achieving higher ranks [12] need to be checked for presence in other sources. Following snapshots show result after comparison between sources.



Zee News : India National Zee News 25 Indian
Zee : India National Zee 25 Indian fishermen ar
Zee News : India National
Zee News
25 Indian fishermen arrested in Sri Lanka
Zee News : India National Zee News 25 Indian
Zee : India National Zee 25 Indian fishermen ar

```
wake of threatens to for but will be wi

d Sri Lanka 25 Indian fishermen arrested
iwal still `agitator` `administrator`: K:
ists try India`s secularism, democracy: !
Inside Compare News


Comparing with BBC
Which is Found : News News in in for Delhi says 2014 The No for says for Delhi under new from to in of Delhi to Day for say


Which is not Found: Zee : India National Zee 25 Indian fishermen arrested Sri Lanka 25 Indian fishermen arrested Sri Lanka
```

Fig 3: Compare News (b)

IV. Results

Implementation part shows that we have successfully compared news for similarity check and recommending missing news. But this can further be improvised for fine results. Most approaches in this area are human based. And very less automated processes are available for comparision. So automated processes can be developed for recommending missing news which can be further tuned to requirements.

V. Conclusion and Future Work

In our system we discussed the design and implementation of Multilevel Ranking for web news information collecting and filtering so that it can be automatically compared with news of its competitors. Besides we introduced concept of removing stop words and inverse document frequency for better comparison with less human intervention. Translation of news in English language to news in language desired by user has been also introduced.

As Future scope we can increase maturity of technique used for similarity check by using a better keyword ranking algorithm.

Also we can recommend entire sentences instead of only keywords. Our system may not only translate news from English to any other language but also it may translate from any language to any other language. Concept based approaches for classification can be implemented with the help of Ontology to handle semantic mismatches.

Acknowledgment

This acknowledgement is intended to be thanks giving gesture to all those people who have been involved directly or indirectly with my dissertation work.

First and foremost, I express my special thanks with gratitude and great respect to my valuable guide Prof. Kavita Kelkar for her keen interest in my project, fruitful suggestions & valuable guidance with motivation & constant encouragement. I am also thankful to her for great patience, constructive criticism & useful suggestion apart from invaluable guidance.

I am also thankful to all the staff members of K.J. Somaiya College of Engineering, Mumbai for their support. I also thank all my friends and family for their continuous support that kept my spirit up during the endeavor.

References

[1] Subrata Saha, Atul Sajjanhar, Shang Gao, Robert Dew, Ying Zaho. "Delivering Categorized News Items Using RSS Feeds and Web Services", IEEE International Conference on Computer and Information Technology (CIT 2010), 2010. p698-703

[2] ZHENG Rui-juan, ZHANG Yang-sen. "Design and Implementation of News Collecting and Filtering System Based on RSS". 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012), 2012. p2295-2298

[3] Hu Haiyan, Su Chang. "Research and Realization based on RSS News Filtering Technology". International Conference on Computer, Mechatronics, Control and Electronic Engineering (CMCE), 2010. p104-106

[4] Shikha Agarwal, Archana Singhal, Punam Bedi. "Classification of RSS News Items Using Ontology", 12th International Conference on Intelligent Systems Design and Applications ISDA, 2012. p491-496

[5] Arnulfo Azcarraga, Michael David Liu, Rudy Setiono. "Keyword Extraction Using Back propagation Neural Networks and Rule Extraction", WCCI 2012 IEEE World Congress on Computational Intelligence June,10-15, 2012.

[6] Ying Qin, "Applying Frequency and Location Information to Keyword Extraction in Single Document", Proceedings of IEEE CCIS2012, 2012. p1398-402.

[7] Rachada Kongkachandra*, Chom Kimpant, Thawatchai Suwanapongt and Kosin Chamnongthai. "Newly-Born Keyword Extraction under Limited Knowledge Resources based on Sentence Similarity Verification", International Symposium on Communications and Information Technologies 2004 (ISCIT 2004) Sapporo, Japan. October 26- 29. 2004. p1183-1187.

[8] Sungjick Lee, Han-joon Kim. "News Keyword Extraction for Topic Tracking", Fourth International Conference on Networked Computing and Advanced Information Management, 2008. p554-559

[9] Fei Liu, Feifan Liu, Yang Liu. "Automatic Keyword Extraction for The Meeting Corpus Using Supervised Approach And Bigram Expansion". Published in SLT 2008. p181-184

[10] R. Mihalcea and P. Tarau. TextRank: "Bringing order into texts". In Proceedings of EMNLP 2004. pp. 404–411. https://github.com/ceteri/textrank

[11] Wenshuo Liu Wenxin Li. "To Determine the Weight in a Weighted Sum Method for Domain-Specific Keyword Extraction" International Conference on Computer Engineering and Technology, 2009. p11-15

[12] "Using Bing Translation API", https://www.microsoft.com/web/post/using-the-free-bing-translation-apis.

Authors

First Author – Maheshkumar Nanda, Student, Masters in Computer Engineering, K. J. Somaiya College of Engineering, Mumbai