A Mini-project Report

on

RSS Feed Classifier

carried out as part of the course Web Technologies and Applications (IT302)

Submitted by

Bhuvan MS (12IT16)
Kartik Koralla (12IT33)
Siddharth Jain (12IT78)
Vinay Rao D (12IT94)
V Sem B.Tech (IT)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

Department of Information Technology

National Institute of Technology Karnataka, Surathkal
November 2014

CERTIFICATE

This is to certify that the project entitled **"RSS Feed Classifier"** is a bonafide work carried out as part of the course **Web Technologies and Applications (IT302)**, under my guidance by Bhuvan MS (12IT16), Kartik Koralla (12IT33), Siddharth Jain (12IT78), Vinay Rao D (12IT94), students of V Sem B.Tech (IT) at the Department of Information Technology, National Institute of Technology Karnataka, Surathkal, during the academic semester V, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Information Technology, at NITK Surathkal.

Place: NITK Surathkal

Date: Signature of the Instructor

DECLARATION

I hereby declare that the project entitled "**RSS Feed Classifier**" submitted as part of the partial course requirements for the course Web Technologies and Applications (IT302) for the award of the degree of Bachelor of Technology in Information Technology at NITK Surathkal during the Jul - Nov 2014 semester has been carried out by us. I declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Further, I declare that I will not share, re-submit or publish the code, idea, framework and/or any publication that may arise out of this work for academic or profit purposes without obtaining the prior written consent of the course Faculty Mentor and Course Instructor.

Signature of the Students:

Name: Bhuvan M S

Name: Kartik Koralla

Name: Siddharth Jain

Name: Vinay Rao D

Place: NITK Surthkal

Date: 17 Nov 2014

ABSTRACT

In today's world, we are witnessing an explosive growth of data on web. With the numbers of users increasing, so is the data and with this growth we need to make internet surfing more efficient i.e. we need to develop proper techniques that give user access to relevant and desired data. Because of so much data on the internet, retrieving the relevant documents is becoming hard and important. Considering the huge number of news readers using internet, here we discuss retrieval of relevant news using RSS (Rich Site Summary or Really Simple Syndication) feeds. In information retrieval, document classification is a necessary requirement and for this many classification algorithms are in use. Algorithm called term frequency-inverse document frequency (TF-IDF) has been used widely for document classification. It classifies documents based on the frequency of terms occurring in the document. This approach has some very elementary limitations. High dimensionality of data is one problem this algorithm faces and also it does not consider the relation among the terms which gives less precision and error prone result (sometimes news contains words in metaphorical context which makes the correlation between words important to classify it to some particular domain). In our approach, we have used an algorithm called weighted concept frequency-inverse document frequency (CF-IDF) with the background knowledge of domain ontology for classification of RSS feed news items. The ontology itself acts as classifier hence removing the need of any trained classifiers. This approach considers the relation between the concept and properties which results in reduction of noise in final output. As we are considering only the key concepts of a domain for classification, this deals with the problem of dimensionality. When the user searches for news about a particular field or domain, CF-IDF aims at returning the most relevant results to that field.

Keywords: RSS feeds, CF-IDF, TF-IDF

TABLE OF CONTENTS

1. Introduction.....1
2. Literature Survey.....3
2.1 Background.....3
2.2 Outcome of Literature Survey.....4
2.3 Problem Statement.....5
2.4 Objectives.....5
3. Methodology.....6
4. Implementation.....9
4.1 Work Done.....9
4.2 Result and Analysis.....10
4.3 Innovative Work.....14
5. Conclusion and Future Work.....15

References.....16

LIST OF FIGURES

Figure 3.1 Proposed Architectural Pattern for classification of RSS feed news items 8
Figure 4.2.1 IPTC News Ontology 11
Figure 4.2.2 Database 13

1. INTRODUCTION

The rate of development today is high and as technology grows, web network also grew rapidly and became huge. Because of this rapid development of web content, there has been a notable increase in the volume of available data. One basic example that can be taken is news. With so much happening all around the world, there is abundant information regarding almost everything from countries to cities to people. This also means that the number of people using web to read news and stay connected has also increased exponentially. And specifically for news, the amount of news in different semi structured formats have also increased recently and hence increasing the need for better classification algorithm. Retrieving relevant information thus has become a challenging task now. It is been seen that current methods which are based on keywords are ineffective and inconsistent i.e. they produce noisy results.

Our reference paper suggests that classification of data based on concepts and utilizing knowledge stored in ontology which is very beneficial Ontology is representation of knowledge pertaining to some specific domain. This method is robust in dealing with semantic mismatches which can improve consistency of our output and it is also used to address scalability problem. Earlier methods required classifier as well but in this method the ontology, which serves as our knowledge base also acts as a classifier. As classifying news based on concept makes more sense than classifying them based on key terms, this method significantly improves the searching result and accuracy.

The proposed method of classification makes use of news items metadata in RSS feeds and tries to strengthen TF-IDF method by modifying it. This method statistically determines the relative importance of terms within a document, in corpus of documents. As mentioned earlier, this method faces elementary problems like scalability (high dimensionality). It also does not consider the relation among the terms which is bad. This affects the precision of classification of news items and leads to inaccurate output. To alleviate the problems discussed above, terms are replaced by key concepts, to handle the problem of dimensionality. We consider only those concepts which are a part of our domain ontology. This ontology follows the International Press Telecomm Council (IPTC) standards of news industry. As the terms are replaced by concepts, this method is called concept frequency-

inverse document frequency (CF-IDF). It considers relations that govern and describe concepts and properties. The ontology itself acts as a classifier so no need for any trained classifiers.

End users are heterogeneous and sometimes they require diverse and different kinds of information. To serve this purpose, a semi structured format called RSS feeds has been developed.RSS basically is a format of delivering regularly changing web content. So whenever the news changes, RSS is updated as well. RSS ensures our privacy and is fast and easy way to stay informed. For particular field of news, we can explain it this way- it "associates news items with feeds". The annotations provided by RSS feeds is generally coarse grained. Finer annotation is provided by our knowledge base ontology. The tags in relation with this feeds do not have a single semantic meaning associated with them and thus they can be interpreted differently. But we need to understand the semantic meaning of news items to properly categorize them so that we can service needs of end user in a better way.

2. LITERATURE SURVEY

2.1 BACKGROUND

First let us discuss about the work that has already been done in the area of document representation and classification.

In late 1980s, most of the clustering or classification algorithms used Vector Space Model (VSM) for document representation. In vector space model, document clustering made use of words to find similarities between documents. Use of Latent Semantic Indexing algorithms (LSI) to improve clustering emerged in 1990. It was in 2002 when actually ontology came into use and improved the quality of document clustering with its concept hierarchy knowledge. This background knowledge was applied during preprocessing state. In 2007, further research was conducted to see usefulness of ontology indexing to combine semantic expressiveness with information retrieval inspired techniques. In 2008 finally an ontology based document clustering technique was developed which uses a domain specific ontology to support the proceeding of document clustering at a conceptual level. In 2009, a semantic similarity based model was developed that made comparisons using WordNet as ontology. It captures semantic similarities among documents that contain semantically similar terms and also assigns a new weight to terms having the semantic relationships among terms that co-occur literally in the document. Later, LSI and ontology were used together. Ontology knowledge served as background for classification and LSI algorithms were applied to reduce the feature vector and improve the performance. Recently even naïve bayse classifier and decision tree method have been used for clustering. Decision tree method, as the name suggests, builds a tree and gives us set of rules. Nearest neighbor approach for short term news recommendation and baysian model for long term have also been tried. Approach based on text indexing systems based on assigning weight to terms produce better results than other systems using some different text representation. The results mostly depend on the weighting scheme we choose. Various methods of finding similarities and the possibilities that can be considered for enhancements in content analysis and text indexing have been discussed[1]. Problems related to information overload on internet have been a primary

concern as it makes web a data rich information poor environment. Solutions telling us how to implement a web based news reader enhanced with a machine learning framework for dynamic content personalization have been implemented. With this, the efficiency of machine learning algorithm for classification of text was examined[2]. Ontological classification has been applied to business intelligence application using ontological classification. The idea was to use ontology as the background knowledge domain and decision tree algorithm[3]. Ontology was beneficial as it also acted as a classifier. All the methods discussed above had some or other limitations which were motivation for developing CF-IDF algorithm rather keep using the traditional approaches like TF-IDF. Limitations are discussed in next section.

2.2OUTCOME OF LITERATURE SURVEY

As we saw in the last section, even decades ago document clustering was an important topic of research and we mentioned some of the techniques used. But there were limitation. Vector Space Model worked but there was still much scope of improving efficiency. With naïve bayse classifier, the problem is that is considers all the terms independent which is one of the major reasons for a noisy result and also it required a trained classifier increasing the overall cost of the method. Properly trained documents are not available very easily or even if they are, the set might just be too small. Decision tree seemed promising but it also had same problem as naïve bayse, requirement of external trained classifier. Also, problems were encountered when multiple class labels were there. Limitations of approach based on keywords have been discussed before. In some of the methods, calculations were too complicated due to usage of all the training examples for classification. When there is no weight difference, performance is solely dependent on the training set. So it can be said that there was need of a method which can handle the problem of scalability, efficiency, accuracy, relationships among the terms in documents. A method based on conceptual representation rather than the traditional representation. What conceptual representation does is it provides text with structured data in the form of annotations, making document management easy. CF-IDF takes care of all this. CF-IDF can be improvised by providing weights. Weight assigned to news items gives more precise result to end user.

2.3 PROBLEM STATEMENT

Classification of RSS news feed items based on RDF ontology using Concept frequency-inverse document frequency algorithm. User can select the category related to which he wants news and after processing the query, system shall return results related to that query.

2.4 OBJECTIVES

1. To eliminate problem of scalability i.e. high dimensionality of data that the current algorithms are facing.
2. Consider relations among different terms in a document to get more accurate result and less noisy output while clustering. (This is one major problem with TF-IDF)
3. Building an RSS feed aggregator using python and putting every news items as titles and news description in the database. (We have used SQLite3)
4. Improving performance of CF-IDF by using weighted CF-IDF to give better precision.

2. METHODOLOGY

The whole implementation revolves around RSS feeds. RSS feeds are basically metadata about abstracts of a website. RSS have 2 elements –

1. Channel Elements (Describes the whole RSS feed)
2. Item Elements (Describes the article the feed is referencing to)

We use RSS feeds to aggregate data in our database.

Earlier TF- IDF was used to classify documents.CF- IDF is almost similar to TF- IDF except that it considers concepts and properties rather than just keywords. Term frequency implies frequency of a term T in some document D. Intuitively, we can say that terms that are rare are more important and informative than the common terms as common terms may appear in documents of different category but if a term is rare, its repetition in documents belonging to different category is improbable. The equation for inverse document frequency is -

$$\mathbf{Idf(T) = \log_2 (d_{num}/d_{freq}(T)) + 1}$$

Here,d_{num} is the total number of documents and similarlyd_{freq}(T) is number of documents with the term T in them. The tf-idf weight of a term is equal to product of its tf weight and its idf weight. So weight is equal to **tf(T, D)** x **idf(T,D).** Weight is directly proportional to the number of times a term occurs in a document and also the rarity of the term in the collection.

We are using ontology as background knowledge for classification. As it has already been discussed, the machine learning algorithms like naïve bayse and support vector machine require trained classifier which is a limitation. We don't need a trained classifier as our ontology not only provides background knowledge but also acts a s a classifier. Moreover, naïve bayse algorithm assumes the terms to be independent which is not the case. It ignores the semantic pattern in data. Latent Semantic Indexing faces with the problem that implicit knowledge represented is difficult to port from one application to another. This is not the case with ontology. Using ontology, it is easy to port knowledge form one application to another. Here precision and recall are taken into account to judge the output, whether its accurate or not. Precision is correctness of classification and recall measures its usefulness. We can check that which factor is important for user and work accordingly. This was basic

introduction to methodology. Now we shall go in depth and discuss the approach and architecture.

To improve the accuracy of classification of documents and relevance in search requires techniques that will consider the semantics of relationships in the data. I our approach we are identifying the key concepts in the preprocessed news items in the feeds, using background domain knowledge. RDF ontology has been designed to act as background knowledge. It takes the semantic relations among the concepts in consideration. To classify, we use TF-IDF algorithm with the key concepts which exist in the domain ontology We assign more weight to concept that appear in title as well as description of the RSS feed news item. This weighted scheme produces more accurate results than the others. All these things altogether can be called as CF-IDF with background knowledge of news domain. We preprocess the news items before classification process to deal with some challenges being faced by other methodologies.

We are using RDF domain ontology as background knowledge. It represents the concepts and the terms related to each concept.

Preprocessing step is basically the aggregation of news feeds from RSS news feeds. After aggregation, we tokenize each news items. If N is a news items, it will be tokenized and will contain T_1, T_2, T_3..... T_n tokens. Words like "the,a,at,on" are called stop words. They enhance the word frequency. Removing the stop words before processing gives better results. We remove semantically meaningless words but otherwise most frequent ones from the text. There are words which are variants of some other base words and have similar semantic interpretations so we consider them as equivalent. To perform this stemming is performed. Stemming means reducing the words to a common form or stem. For example bored, boring, bore all can be reduced to bor.

After this, we need to identify the concepts and sub concepts using background knowledge from the domain ontology. Ontology concepts are traversed against concepts of news itemsto check what concepts are found in title and description of RSS feed news items. Now CF-IDF is performed to find the relative importance of concepts. **CF-IDF=CF** x **IDF** where CF is occurrence of concept c_i in document d_j and IDF is occurrence of c_j in a set of document D. While assigning weights, we keep in mind that concepts appearing in both title and description are given more weight than the concepts coming either in title or description. The news items are then annotated using attributes of ontology and are stored, so that the future

queries can be executed faster without going through the preprocessing step. If some category c is in title and description, its weight will be more than some other category l, which is only in title or description. This is the whole approach for classification of news feeds. Coming to architecture, we have implemented a system of RSS feed news items classification. The figure 1 below shows the system architecture and how the process works.

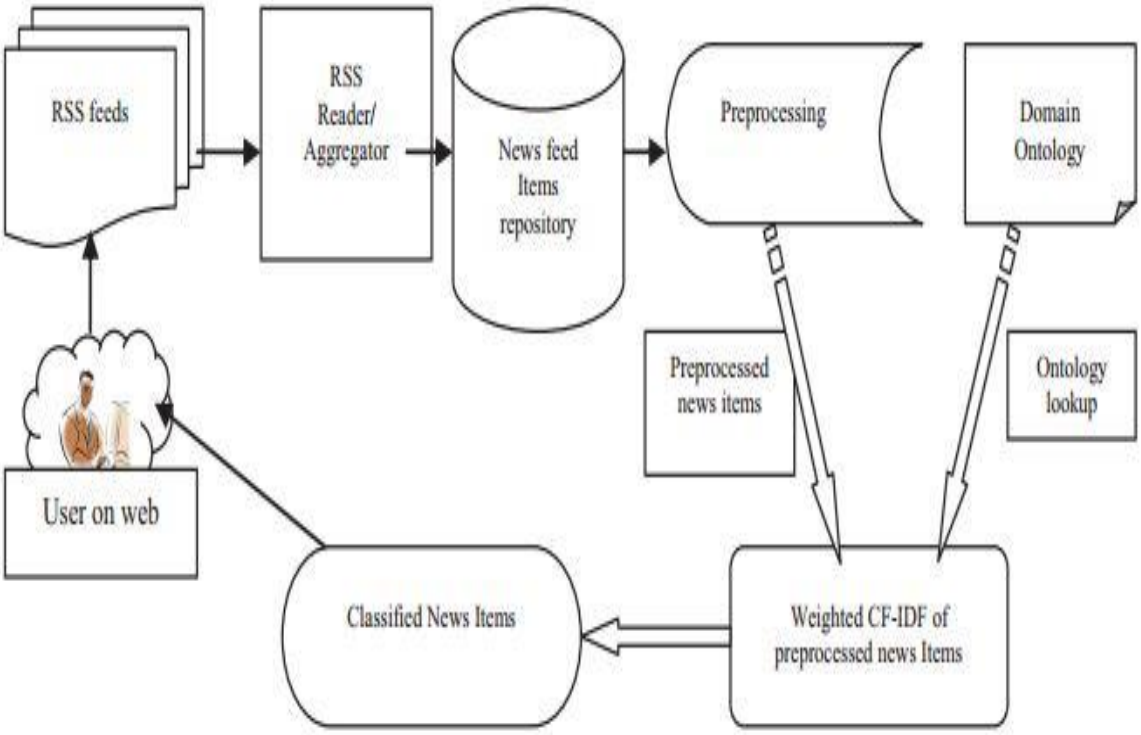

Figure 3.1: Proposed Architectural Pattern for classification of RSS feed news items

4. IMPLEMENTATION

4.1 WORK DONE

We used software components and open source tools to implement the method. The operating system used for development is Windows 7. SQLite3 is used for database at backend. The aggregator code has been written in Python 2.7 and flask is the micro frame work for web application which we have used.

Our system will have following features:

1. Links are fed to aggregator for building our data base. The news items are then classified into different categories using Ontology and CF-IDF algorithm. RSS links from Times of India etc. were used to create an authentic data base.
2. User can select the category of news he/she wants to read.
3. System responds with news belonging to that category.

The feed aggregator written in python uses inbuilt libraries and functions to create tables and query the database by importing sqlite3.HTML parser strips the html tags and keeps the important data intact. The RSS links are stored in a text file and the aggregator's function is to put the corresponding news items in our database. SQLite3 is a software library that implements a self contained, transactional SQL database engine. Using this we store the news items as title and description. With help of Protégé IDE, ontology matching IPTC standards was used as the background domain knowledge. Using the ontology, we get an excel file containing concepts and their subclass. This excel file is used to build a dictionary where keys are the concepts mapping to all the subclasses (as a list) for the corresponding concept. These dictionaries are stored in a text file using pickle (in python). Just to expand our domain of categories, referencing (SITES USED BY KARTIK HERE) we made an extra text file containing additional words related to concepts so that category matching during the processing is more accurate. These are also converted into dictionary and appended to the dictionaries developed from the ontology. Now, to perform weighted CF-IDF, we use this dictionary (retrieved using pickle) and give weight to concepts according to their occurrences in title and description to decide the category. Flask, which is a python based web framework was used to meet the front end requirements.

4.2 RESULT AND ANALYSIS

Feed aggregator parses all the news feeds from the links and places in the sqlite3 database. Some of the important RSS Feeds URL used to aggregate the news feeds are listed below:

http://feeds.bbci.co.uk/news/business/rss.xml

http://feeds.bbci.co.uk/news/education/rss.xml?edition=uk

http://feeds.bbci.co.uk/news/health/rss.xml?edition=uk

http://feeds.bbci.co.uk/news/science_and_environment/rss.xml?edition=uk

http://feeds.bbci.co.uk/news/politics/rss.xml?edition=uk

http://feeds.bbci.co.uk/news/entertainment_and_arts/rss.xml

http://www.fbi.gov/news/news_blog/rss.xml

http://www.crimesolutions.gov/feed.svc/Fetch/Rss?61d06558-b6ee-41b5-b02c-cc8afb65e3e0

http://www.thehindu.com/sci-tech/?service=rss

http://feeds.hindustantimes.com/HT-Fashion

http://feeds.hindustantimes.com/HT-Travel

http://feeds.hindustantimes.com/HT-Books

http://feeds.hindustantimes.com/HT-brunch-topstories

http://ibnlive.in.com/ibnrss/rss/buzz/buzz.xml

http://ibnlive.in.com/ibnrss/rss/trends/society.xml

http://www.globalissues.org/news/topic/165

http://www.globalissues.org/news/topic/166

http://www.pewforum.org/feed//

http://billtammeus.typepad.com/my_weblog/atom.xml

http://topics.nytimes.com/top/reference/timestopics/subjects/r/religion_and_belief/index.html?rss=1

http://feeds.bbci.co.uk/news/world-middle-east-17258397/rss.xml

http://www.economist.com/topics/war-and-conflict/index.xml

http://www.economist.com/topics/war-and-conflict/rss

http://www.globalissues.org/issue/178/climate-change-and-global-warming

http://www.globalissues.org/news/topic/137

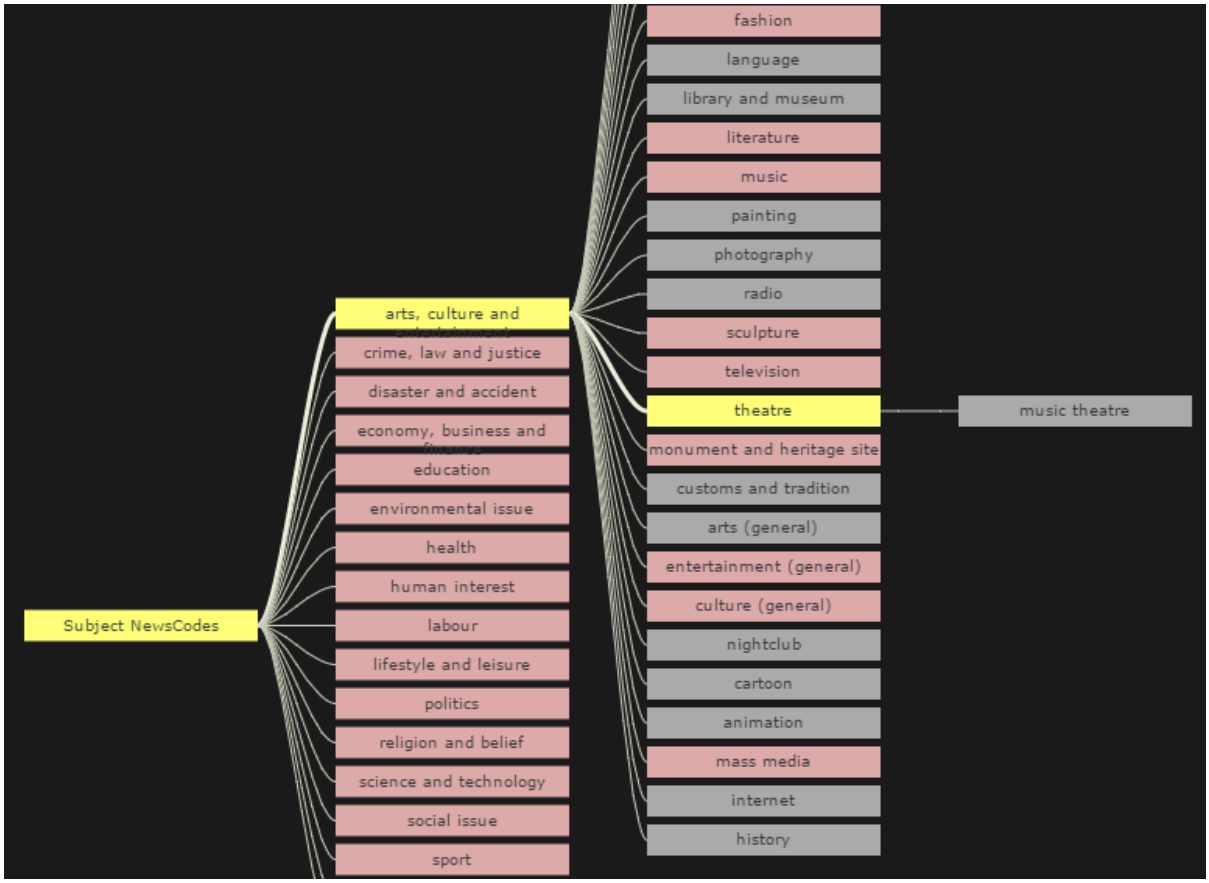

Fig.4.2.1: IPTC News Ontology

News RDF Ontology acquired from IPTC site, contains all concepts mapped to their respective terms as shown in the fig.4.2.1. These concepts have to be mapped to news items, by matching the title and description of the news item, and calculating the weighted CF-IDF score. Additional bag of words with respect each category is used for better accuracy of classification.

Feed Categorizer first removes all the stop words like ‘the’, ‘and’, ‘is’ etc., that occur in the news item as these words does determine the category to which the feed may belong to. Also it proved to be efficient to stem the words of both conceptual terms from ontology as well as news items before matching, as it helps to catch all the morphological forms of words. Natural Language Tool Kit in python provides Porter Stemmer which was used to stem the text before calculating CF-IDF scores. These pre-processing steps considerably increased the number of CF-IDF scores for news items. The concept for which the news item has maximum CF-IDF score is the concept that the news-item emphasises hence it can be decided that the news item belongs to that category.

This naive approach can be improved by applying weighted CF-IDF. Intuitively, the concept which has greater CF_IDF scores in both title and description is more relevant to that news rather than a concept that scores very high in title but very low in description or vice-versa. Hence it adding the CF-IDF scores for title and CF-IDF score of description of a single news item is considered to be aggregated/weighted CF-IDF score which determines the category of the news item. In addition, based on experimentation certain weights can be decided that could be assigned to title and description for more accurate aggregation of CF-IDF scores. For example, 1.25 may be wight for title and 1.0 can be weight for description, which implies CF-IDF score for the title has more effect on the category of the news item than the description. Such weights can be determined by machine learning methods like linear regression in future, but we assume the weights to be equal for title and description for initial classification of news feeds.

The following are number of news feeds that were categorized under each category:
Out of 1466 news feeds:
"arts culture and entertainment"=>"89"
"crime law and justice"=>"75"
"disaster and accident"=>"13"
"economy business and finance"=>"250"
"education"=>"43"
"environmental issue"=>"34"
"general"=>"26"
"health"=>"113"
"human interest"=>"67"
"labour"=>"106"
"lifestyle and leisure"=>"61"
"politics"=>"94"
"religion and belief"=>"41"
"science and technology"=>"27"
"social issue"=>"83"
"sport"=>"286"
"unrest conflicts and war"=>"29"
"weather"=>"29"

This process can be dynamic as news items are gathered from RSS Feeds which dynamical are updated by the news site administrators. Periodically running the Feed Aggregator and Feed Categorizer updates the database where all news items are stored along with the category they belong to. This makes the whole system dynamic. Here we use unsupervised classification where we do not have training data, hence for accuracy of classification an additional bag of words has been used, where we observed that comparatively more feeds were categorized with higher CF-IDF scores than those when additional bag of words were not used. By comparing the results of normal categorizer without bag of words and the categorizer with bag of words we got 63.20% match. Since this is unsupervised classification we cannot for sure say which categorizer is correct. Perhaps we put up a hypothesis that categorizer with bag of words would be a bit less accurate but covers a larger domain of news feeds. On further experimenting to get training data we might be able to distinguish these methodologies based on their accuracy, which would be future work which could follow this project.

This database synthesized after classification acts as access layer which is accessed by Flask code which connects to database to pull the data by sql queries and display it on web pages where the client can browse through news categories.

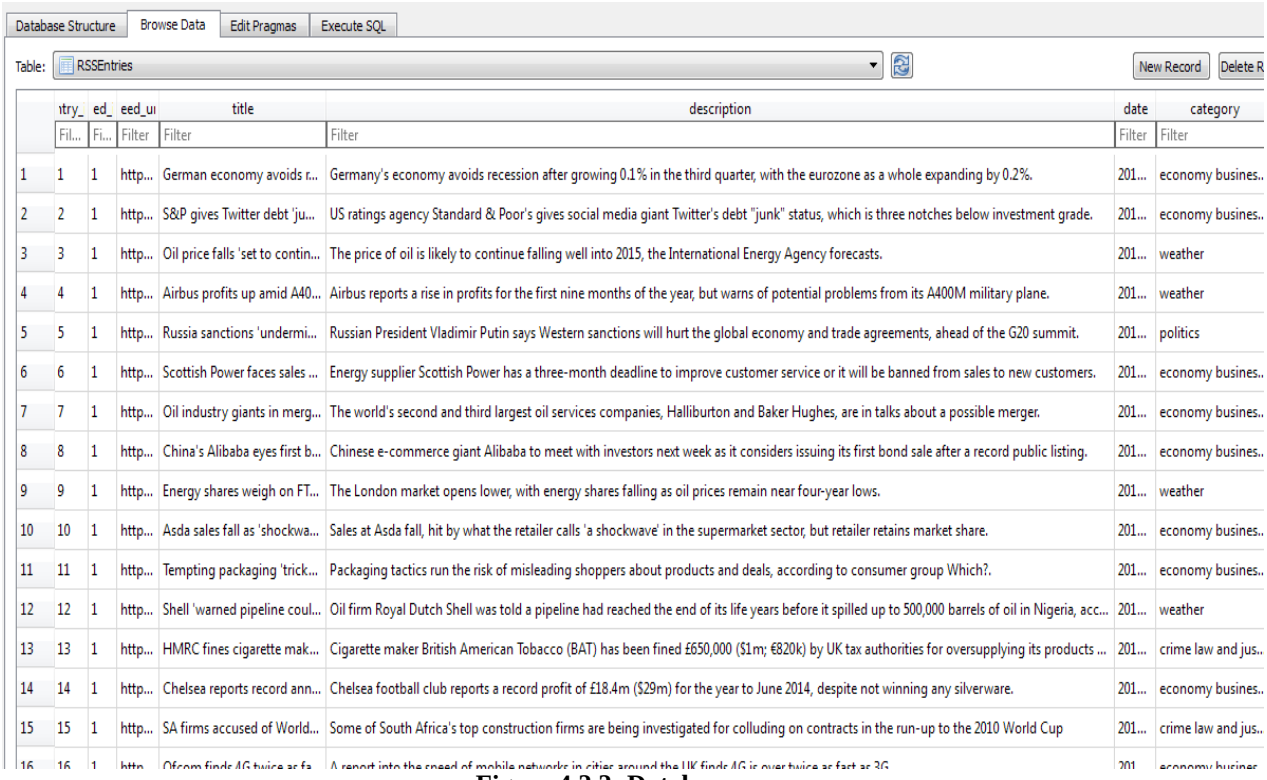
Database Structure | Browse Data | Edit Pragmas | Execute SQL

Table: RSSEntries — New Record | Delete R

	ıtry_	ed_	eed_ur	title	description	date	category
	Fil...	Fi...	Filter	Filter	Filter	Filter	Filter
1	1	1	http...	German economy avoids r...	Germany's economy avoids recession after growing 0.1% in the third quarter, with the eurozone as a whole expanding by 0.2%.	201...	economy busines..
2	2	1	http...	S&P gives Twitter debt 'ju...	US ratings agency Standard & Poor's gives social media giant Twitter's debt "junk" status, which is three notches below investment grade.	201...	economy busines..
3	3	1	http...	Oil price falls 'set to contin...	The price of oil is likely to continue falling well into 2015, the International Energy Agency forecasts.	201...	weather
4	4	1	http...	Airbus profits up amid A40...	Airbus reports a rise in profits for the first nine months of the year, but warns of potential problems from its A400M military plane.	201...	weather
5	5	1	http...	Russia sanctions 'undermi...	Russian President Vladimir Putin says Western sanctions will hurt the global economy and trade agreements, ahead of the G20 summit.	201...	politics
6	6	1	http...	Scottish Power faces sales ...	Energy supplier Scottish Power has a three-month deadline to improve customer service or it will be banned from sales to new customers.	201...	economy busines..
7	7	1	http...	Oil industry giants in merg...	The world's second and third largest oil services companies, Halliburton and Baker Hughes, are in talks about a possible merger.	201...	economy busines..
8	8	1	http...	China's Alibaba eyes first b...	Chinese e-commerce giant Alibaba to meet with investors next week as it considers issuing its first bond sale after a record public listing.	201...	economy busines..
9	9	1	http...	Energy shares weigh on FT...	The London market opens lower, with energy shares falling as oil prices remain near four-year lows.	201...	weather
10	10	1	http...	Asda sales fall as 'shockwa...	Sales at Asda fall, hit by what the retailer calls 'a shockwave' in the supermarket sector, but retailer retains market share.	201...	economy busines..
11	11	1	http...	Tempting packaging 'trick...	Packaging tactics run the risk of misleading shoppers about products and deals, according to consumer group Which?.	201...	economy busines..
12	12	1	http...	Shell 'warned pipeline coul...	Oil firm Royal Dutch Shell was told a pipeline had reached the end of its life years before it spilled up to 500,000 barrels of oil in Nigeria, acc...	201...	weather
13	13	1	http...	HMRC fines cigarette mak...	Cigarette maker British American Tobacco (BAT) has been fined £650,000 ($1m; €820k) by UK tax authorities for oversupplying its products ...	201...	crime law and jus..
14	14	1	http...	Chelsea reports record ann...	Chelsea football club reports a record profit of £18.4m ($29m) for the year to June 2014, despite not winning any silverware.	201...	economy busines..
15	15	1	http...	SA firms accused of World...	Some of South Africa's top construction firms are being investigated for colluding on contracts in the run-up to the 2010 World Cup	201...	crime law and jus..
16	16	1	http...	Ofcom finds 4G twice as fa...	A report into the speed of mobile networks in cities around the UK finds 4G is over twice as fast as 3G	201...	economy busines..

Figure 4.2.2: Database

4.3 INNOVATIVE WORK

We have used additional terms for each concept apart from the terms found in ontology to expand the concept mapping to cover larger domain of news feeds. Classification used in the referencing paper was supervised approach where training data was available. In our implementation we classified over unsupervised data by using both news ontology and bag of words to classify the news. We have compared the difference in classification of news items done using news ontology only and using both news ontology and bag of words approach.

5. CONCLUSION AND FUTURE WORK

RSS Feeds technology has enabled News organizations to produce News Feeds and update regularly. This is more convenient way to read news as users can just load new news items after feed update without loading the whole website. Hence the practice of reading news through news feeds has become more popular. This has lead to increase in the number of news feeds and hence a powerful news classifier is needed to organize the news items for the convenience of the user.

Over the years, TF-IDF, bag of words approach, and other similar methods were used for news classification, but the results achieved were not up to mark. Hence using CF-IDF with News Ontology of IPTC standard helped mapping of news items to specific concepts and in turn CF-IDF scores classified the news items with improved accuracy as mentioned in the referenced paper. The hypothesis of weighted CF-IDF by assigning weights to CF-IDF scores of title and description proved to be correct by showing further improved accuracy. In addition to news ontology using bag of words approach and classifying using a combined approach helps to increase the domain of news feeds as bag of words would classify news items which have terms not represented in the IPTC news ontology, but accuracy may vary.

Future work includes assigning varying weights to title and description and experimenting to get optimal weights by machine learning methodologies. Using supervised learning methods accuracy comparison needs to be done the two approaches we have compared, one in which just Ontology concepts are mapped, another in which Ontology concepts and additional bag of words have been mapped to news items.

The news feed data is streaming and its number is increasing tremendously, hence there is need for building scalable models using distributed cloud computing frameworks like Apache Storm which can handle streaming data through topologies, and Apache Spark which can support streaming data as well as scalable machine learning on large data.

REFERENCES

[1]. Gerard Salton and Chris Buckley (1987) “Term Weighting Approaches in Automatic Text Retrieval”, 87-881

[2]. Ioannis Katakis, Grigorios Tsoumakas, Evangelos Banos, Nick Bassiliades and Ioannis Vlahavas (2008) “An Adaptive Personalized News Dissemination System”

[3]. A. Martin, D. Maladhy and Dr. V. Prasanna Venkatesan (2011) “A Framework For Business Intelligence Application Using Ontological Classification”

[4]. IPTC, International Press Telecommunication Council News Ontology, 2010, Available: http://www.iptc.org/site/Home/, Visited on 27^{h} October 2014

[5]. Armin Ronacher, Flask web development one drop at time,published year, Avalable: http://flask.pocoo.org/, visited: Visited on 2^{h} November 2014

[6]. http://dictionary.cambridge.org/

[7]. http://www.myvocabulary.com/

[8]. http://www.macmillandictionary.com/

[9]. http://www.enchantedlearning.com

[10]. http://www.whatisrss.com/