Mining Diverse *Views* from Related Articles

Ravali Pochampally
Center for Data Engineering
IIIT Hyderabad
Hyderabad, India
ravali@research.iiit.ac.in

Kamalakar Karlapalem
Center for Data Engineering
IIIT Hyderabad
Hyderabad, India
kamal@iiit.ac.in

ABSTRACT
The world wide web allows for diverse articles to be available on a news event, product or any topic. It is not impossible to find a few hundred articles that discuss a specific topic thus making it difficult for a user to quickly process the information. Summarization condenses huge volume of information related to a topic but does not provide a delineation of the issues pertaining to it. We want to extract the diverse issues pertaining to a topic by mining views from a collection of articles related to it. A view is a set of sentences, related in content, that address an issue relevant to a topic. We present a framework for extraction and ranking of views and have conducted experiments to evaluate the framework.

Categories and Subject Descriptors
H.5 [**Information Systems**]: Information Interfaces and Presentation

General Terms
Human Factors, Experimentation

Keywords
text mining, views, diversity, information retrieval

1. INTRODUCTION
The world wide web is a storehouse of information. Users who want to comprehend the content of a particular topic (e.g. FIFA 2010) are often overwhelmed by the volume of text available on the web. Websites which organize information based on content (google news[1]) and/or user ratings (amazon[2], imdb[3]) also output several pages of text in response to a query. It is difficult for an end-user to process all the text presented.

Multi-Document Summarization [2] is a prominent Information Retrieval (IR) technique to deal with this problem of *information overload*. But summaries typically lack the semantic grouping to present the multiple views addressed by a group of articles. Providing diverse views and allowing users to browse through them will faciliate the goal of information exploration by providing the user a definite and detailed snapshot of their topic of interest.

[1]http://news.google.com/
[2]http://www.amazon.com/
[3]http://www.imdb.com/

Articles which pertain to a common topic (e.g swine-flu in India) are termed as 'related'. By isolating views we aim to organize content in a detailed manner than that of summarization. We define a *view* as

A sentence or a set of sentences which broadly relate to an issue addressed by a collection of related articles and aid in elaborating the different aspects of that issue

1.1 Motivating Example
Here is a pair of views obtained by our framework. Both the views are mined from Dataset 1. The number in the curly brackets indicates the ID of the article from which the sentence is extracted. Description of datasets is given in Table 1.

Example Views

1. The irresponsibility of the financial elite and US administrations has led the US economy to the brink of collapse. {18} On Friday, the Dow was down a mere 0.3% on the week - but to get there, the Fed and the Treasury had to pump hundreds of billions into the global financial system. {14} The collapse of the oldest investment bank in the country could strongly undermine the whole US financial system and increase the credit crisis. {3} After a week that saw the collapse of Lehman Brothers, the bailout of the insurer AIG and the fire sale of Merrill Lynch and the British bank HBOS, policy makers hit back, orchestrating a huge plan to sustain credit markets and banning short sales of stock. {48} It was a dramatic reversal from the first half of the week, when credit markets virtually seized up and stocks around the globe plunged amid mounting fears for the health of the financial system. {18}

2. The Swiss National Bank is to pump USD 27 billion into markets and the Bank of Japan (BOJ) valued its part in the currency swap with the Federal Reserve at 60 billion. {35} The Bank of Canada was also involved, and The Bank of England said it would flood 40 billion into the markets. {26} And, despite the agreements that Barclays Capital and Bank of America will sign with executives at Lehman Brothers or Merrill Lynch, it is the hunting season in the banking world for the crème de la crème. {14}

The first view details the breakdown of the US economy along with a few signs of damage control. The second view reports the actions of various banks during the financial turmoil in 2008. These views capture a glimpse of the specific issues pertaining to the topic of 'financial meltdown'. A list of such diverse views would organize the content of a collection of related articles and provide a perspective into that collection.

The problem statement is

Given a corpus of related articles A, identify the set V of views pertaining to A, rank V and detect the most relevant view (MRV) along with the set of outlier views (OV)

1.2 Related Work

Allison et. al [1] [8] proposed that providing multiple view-points of a document collection and allowing to move among these view-points will facilitate the location of useful documents. Representations, processes and frameworks required for developing multiple view-points were put forth.

Tombros et al. [10] proposed the clustering of Top-Ranking Sentences (TRS) for efficient information access. Clustering and summarization were combined in a novel way to generate a personalized information space. Clusters of TRS were generated by a hierarchical clustering algorithm using the group-average-link method. It was argued that TRS clustering presents better information access than routine document clustering.

TextTiling [5] is a technique for subdividing text into multi-paragraph units that represent passages or subtopics. It makes use of patterns of lexical co-occurence and distribution. The algorithm has three parts: tokenization into sentence-sized units, determination of a score for each unit and detection of sub-topic boundaries. Sub-topic boundaries are assumed to occur at the largest valleys in the graph that result from plotting sentence-units against scores.

1.2.1 Views vs. Summary

Summary and views generated for Dataset 5 are here - (https://sites.google.com/site/diverseviews/comparison)
The summary is generated by update summarization 'baseline algorithm' [6]. It is conspicuous by the lack of organization. Though successful in covering the salient features of the review dataset, it groups several conflicting sentences together (observe the last two sentences of the summary). The views generated by our framework present an organized representation by generating clusters of semantically related sentences. As is evident, the first view is discussing the positive attributes of hotel taj krishna in hyderabad while the second view is negative in tone. The third and fourth views discuss specific aspects of the hotel such as the food and the facilities available. Presenting multiple views for a topic allows us to model the diversity in its content. Our representation is concise as the average number of sentences per view was found to be 3.9. In our framework, we address two drawbacks of summarization - lack of organization and verbosity (due to user-specified parameters).

ID	Source	Search Term	# Articles
1	google news	financial meltdown	49
2	google news	swine flu india	100
3	google news	israel attacks gaza	24
4	amazon.com	the lost symbol	25
5	tripadvisor.com	hotel taj krishna	20
6	tripadvisor.com	hotel marriott	16
7	google news	fifa vuvuzela	39
8	google news	gulf oil spill	26

Table 1: Datasets

1.3 Contributions

The main contributions of this work are

1. Defining the concept of a *view* over a corpus of related articles
2. Presenting a framework for mining diverse views
3. Ranking the views based on a quality parameter (*cohesion*) defined by us and
4. Presenting results to validate the framework

1.4 Organization

In section 2, we elaborate on the framework for the extraction of views. MRV, OV and the ranking mechanism are explained in detail in section 2.5. Section 3 is for experimental evaluation and discussion. In section 4, we sum up our contributions and outline the future work.

2. EXTRACTION OF VIEWS

In this section, we detail the steps involved in the extraction of views and define a quality parameter for ranking the views according to their relevance. Figure 1 presents an overview of the framework by depicting the steps involved in the algorithm. Input and output are specified for each step of the algorithm.

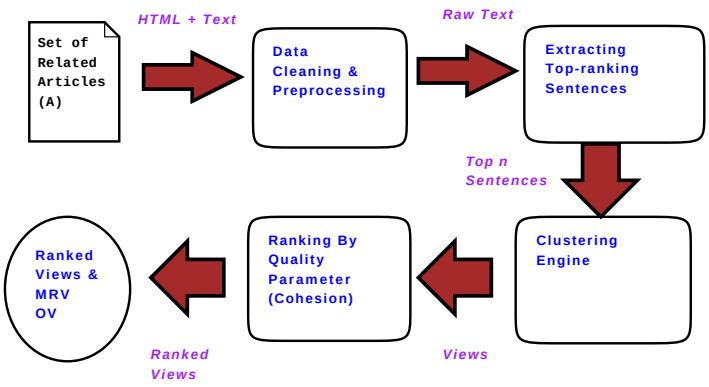

Figure 1: Framework

2.1 Datasets

Articles which make relevant points about a common topic but score low on pairwise cosine similarity can be included in our datasets because we aim to present multiple views from a set of related articles, rather than group them based on overall content similarity. We used data from news aggregator and review web sites as they group articles discussing a common topic, inspite of the low semantic similarity between them. We crawled articles published between a range of dates when the activity pertaining to a relevant topic peaked. For example, we crawled articles published on 'gulf oil spill' between 15 April 2010 and 15 July 2010 when the news activity pertaining to that topic was maximum. We crawled websites which provided rss feeds or had a static html format that could be parsed. Table 1 provides the description of datasets. For instance, Dataset 1 is collected from google news using the search term 'financial meltdown' and contains 49 articles. Datasets can be found here - (https://sites.google.com/site/diverseviews/datasets)

2.2 Data Cleaning and Preprocessing

Web data was collected using Jobo[4], a java crawler. The data was given as an input to the data cleaning and preprocessing stage. Data Cleaning is important as it parses the html data and removes duplicates from the articles. We define a 'duplicate' as an article having the exact syntactic terms and sequences, with or without the formatting differences. Hence, by our definition, duplicates have a cosine similarity value of one.

Text data devoid of html tags is given as an input to the data preprocessing stage. Stemming and stopword removal are performed in the preprocessing stage. Stemming is the process of reducing inflected (or derived) words to their stem or root form. (example: running to run, parks to park etc.) In most cases, these morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications. Stopwords are the highly frequent words in english language (example: a, an, the, etc.). Owing to their high frequency, and usage as conjunctions and prepositions, they do not add any significant meaning to the content. Hence, their removal is essential to remove superfluous content and retain the essence of the article. In order to capture the user notion, the review datasets were not checked for typographical and grammatical errors and were retained verbatim. Python modules HTMLParser[5] and nltk.wordnet[6] were used to parse the html data and perform stemming respectively. IR metrics such as word frequency and TF-IDF[7] were extracted for future analysis.

2.3 Extraction of Top-Ranking Sentences

A dataset consisting of many articles and having content spanning various issues needs an pruning mechanism to extract sentences from which the views can be generated. We prune a dataset by scoring each sentence in it and extracting the top-ranked ones. A list of notations used in our discussion is given in Table 2

$T_{i,j}$	:	$tf - idf_{i,j}$
$tf - idf_{i,j}$	:	TF-IDF of term t_i in article d_j
$tf - idf_{i,j}$	:	$tf_{i,j} * idf_i$
$tf_{i,j}$	:	$\frac{n_{i,j}}{\sum_k n_{k,j}}$
$n_{i,j}$	:	Number of occurences of t_i in article d_j
$\sum_k n_{k,j}$	:	$\sum$ of occurences $\forall$ t_k in article d_j
idf_i	:	$\log \frac{\|D\|}{\|d : t_i \in d\|}$
$\|D\|$	:	Total number of articles in the corpus
$\|d : t_i \in d\|$	:	Number of articles which have the term t_i

Table 2: Notations

Let $< S_1, S_2, S_3 ... S_n >$ be the set of sentences in an article collection. $tf - idf_{i,j}$ (TF-IDF) of a term t_i in article d_j is obtained by multiplying its weighted term frequency $t_{i,j}$ and inverse document frequency idf_i. A high value of $tf - idf_{i,j}$ ($T_{i,j}$) is attained by a term t_i which has a high frequency in a given article d_j and low occurence rate among the spectrum of articles present in that collection. Appearance of some words in an article is more indicative of the issues addressed by it than others. $T_{i,j}$ is a re-weighting of word importance, though it increases proportionally by the number of times a word appears in an article, it is offset by the frequency of the word in the corpus. We consider a product of the $T_{i,j}$ of constituent words in a sentence to be a good indicator of its significance. A product can be biased by the number of words in a sentence hence, we normalize the product by dividing it with the length of the sentence. Given the notation above, we thus define the importance I_k, of a sentence S_k, belonging to an article d_j and having r constituent words as

$$I_k = \frac{\prod_{i=1}^{r} T_{i,j}}{r}$$

$T_{i,j} = tf - idf$ of word $w_i \in S_k \wedge d_j$
$I_k =$ product of $tf - idf$ of $\forall w_i$ normalized according to sentence (S_k) length, r

Logarithm normalization was not used as the σ value for r was 2.2 and variance in its value was not exponential. Sentences are arranged in the non-increasing order of their *importance* (I) scores. We choose the top n sentences for our analysis. Experiments are conducted to correlate the range of n with the corresponding score obtained by our ranking parameter.

2.4 Mining Diverse Views

A measure of similarity between two sentences is required to extract semantically related views from them. Semantic similarity calculates the correlation score between sentences based on the likeness of their meaning. Mihalcea et al. [7] proposed that the *specificity* of a word can be determined using its inverse document frequency (idf). Using a metric for word to word similarity and specificity, the semantic similarity of two text sentences S_i and S_j , where w represents a word in a sentence, is defined by them as

[4]http://java-source.net/open-source/crawlers/jobo
[5]http://docs.python.org/library/htmlparser.html
[6]http://www.opendocs.net/nltk/0.9.5/api/nltk.wordnet-module.html
[7]http://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html

$$sim(S_i, S_j) = \frac{1}{2}\left(\frac{\sum_{w \in \{S_i\}}(maxSim(w, S_j) * idf(w))}{\sum_{w \in \{S_i\}} idf(w)} + \frac{\sum_{w \in \{S_j\}}(maxSim(w, S_i) * idf(w))}{\sum_{w \in \{S_j\}} idf(w)}\right)$$

This metric is used for our analysis as it combines the semantic similarities of each text segment with respect to the other. For each word w in the segment S_i, we identify the word in segment S_j that has the highest semantic similarity, i.e. $maxSim(w, S_j)$, according to some pre-defined word-to-word similarity measures. Next, the same process is applied to determine the most similar word in S_j with respect to the words in S_i. The word similarities are then weighed with corresponding word specificities, summed up and normalized according to the length of each sentence.

Wordnet based similarity measures score well in recognizing semantic relatedness [7]. Pucher [9] has carried out the performance evaluation of all the wordnet based semantic similarity measures and found that *wup* [4] is one of the top performers in capturing semantic relatedness. We also chose *wup* because it is based on the path length between synsets of words and its performance is consistent across various parts-of-speech (POS). We used Python nltk.corpus[8] to implement wup. Pairwise semantic similarity $sim(S_i, S_j)$ or $s_{i,j}$ is a symmetric relation. Thus, we used the upper traingle of the similarity matrix (X) to reduce computational overhead.

$$\forall s_{i,j} \in X \implies \{s_{i,j} = s_{j,i}\}$$

We used clustering to proceed from a set of sentences to views containing similar content. The similarity-matrix (X) was given as an input to Python scipy-cluster[9] which uses Hierarchical Agglomerative Clustering (HAC). HAC was used because we can terminate the clustering when the values of the scoring parameter converge without explicitly specifying the number of clusters to output.

Each cluster comprises of sentences grouped according to the similarity measure ($s_{i,j}$) discussed above. Hence, it is logical to treat them as views discussing a specific issue. In the next section, we propose a quality parameter for the ranking and evaluation of views.

2.5 Ranking of Views

Qualitative parameter for ranking the views focuses on average pairwise similarity between constituent sentences of a view V in order to define its cohesion (C). We define cohesion as

$$C = \frac{\sum_{i,j \in V} s_{i,j}}{len(V)}$$

$$s_{i,j} = sim(T_i, T_j)$$

$V =$ set of sentences (T_i) comprising the view

$len(V) =$ number of sentences in the view.

[8]http://nltk.googlecode.com/svn/trunk/doc/api/nltk.corpus-module.html

[9]http://code.google.com/p/scipy-cluster

As per our definition, higher the value of cohesion, greater is the content similarity between the sentences of a view. Our framework wanted to ascribe importance to views with maximum pairwise semantic similarity. Thus, we defined Most Relevant View (MRV) as the view with maximum value of cohesion, i.e., maximum content overlap amongst its constituent sentences. Outlier views (OV) represent the set of views containing a single sentence. They are termed as outliers because their semantic similarity with others is too low to have any meaningful grouping. We rank all the views in the non-increasing order of their cohesion. As their corresponding pair-wise similarity is zero, outlier views have a cohesion value of zero. Hence, we order outlier views according to their importance (I) scores.

2.6 Framework for Extracting Views

Algorithm 1 provides the steps involved in mining diverse views from a set of related articles. The articles are cleaned by parsing the html and removing duplicates. IR metrics such as TF-IDF are collected before calculating the importance (I) of each sentence. The sentences are ranked in the non-increasing order of their importance to pick the top n sentences. We calculate the pair-wise semantic similarity between the chosen sentences to cluster them. Clustering is used to generate semantically related views from a set of disparate sentences. We rank the views according to the quality parameter proposed by us.

3. EXPERIMENTAL EVALUATION

Extraction of Top-Ranking sentences requires the number of constituent sentences (n) as an input. The ideal range of values for an input parameter is the one which can maximize the cohesion of views and determining it is a critical part of our framework. Hence, we analysed the result data to find the relevant range for n.

An input parameter producing views where the median cohesion is greater than (or equal to) the mean is preferred. As the mean is influenced by the outliers in a dataset, the median being at least as high as the mean indicates consistency across the values of cohesion. If all the values of mean cohesion are greater than that of median, the input parameter yielding views with the maximum mean cohesion is preferred.

We collected statistics about the cohesion (mean, median), number of views, outliers etc. for values of n equal to 20, 25, 30, 35, 40, and 50. The results are presented in Table 5. ID indicates the dataset-ID (as per Table 1), TRS stands for the number of Top-Ranking sentences, V and O stand for the number of views and outliers respectively.

Figures 2 to 9 plot the variation in the mean and median cohesion in relation to the number of TRS (n). The value of n is plotted on the horizontal axis and the value of cohesion is plotted on the vertical axis. We can deduce from the graphs that the mean and median cohesion are peaking for $20 \leq n \leq 35$. The exact breakup of the value of n yielding the best cohesion for all the datasets is provided in Table 3.

As evident from our results, choosing more top-ranking sentences need not necessarily lead to views with better cohesion. To extract views with best cohesion one can start with

Algorithm 1 Mining Diverse Views

Require: Related Articles A
Ensure: Ranked Views V with MRV and OV

for all a in A **do**
    aClean ← ParseHTML(a)

    if aClean is not *duplicate* **then**
        ACLEAN ← ACLEAN + aClean
    else
        discard aClean
    end if
end for
for all a in ACLEAN **do**
    a ← removeStopwords(a) //ranks.NL stopwords
    ASTEM ← ASTEM + stem(a) //nltk stemmer
end for
for all a in ASTEM **do**

    for all *word* in a **do**
        computeTFIDF(*word*)
    end for
end for
for all *sentence* in ASTEM **do**
    rankedSentences ← calculateImportance(*sentence*) //section 2.3
end for
topN ← pickTOPsentences(rankedSentences,n) //as per importance (I)
for all *sentence*1 as $s1$ in topN **do**
    for all *sentence*2 as $s2$ in topN **do**
        if ($s1$,$s2$) not in simMatrix **then**
            simMatrix ← simMatrix + calculateSimilarity($s1$,$s2$)
        end if
    end for
end for
rawViews←clusteringEngine(simMatrix)//scipy-cluster
for all *view* in rawViews **do**
    views←views+calculateCohesion(*view*)//section 2.5
end for
rankedViews ← rankByCohesion(views)
MRV ← chooseMaxCohesion(rankedViews)
OV ← chooseZeroCohesion(rankedViews)

a lower bound (e.g. 20) of top-ranking sentences and incrementally add x sentences until one reaches an upper bound (e.g. 35). Incremental clustering [3] can be used to obtain views. The cohesion values can be compared to present the set of views which yield the best cohesion. Below we present three views mined by our framework. The value of n for each view is the one which yields best cohesion for that dataset (as presented in Table 3)

Example 1 | fifa vuvuzela (7) | n: 35 | cohesion: 40.71 (MRV)
The true origin of the vuvuzela is disputed, but Mkhondo and others say the tradition dates back centuries - "to our forefathers" - and involves the kudu.{5} The plastic trumpets, which can produce noise levels in excess of 140 decibels, have become the defining symbol of the 2010 World Cup. {12} For this reason, there is no doubt that the vuvuzela will become one of the legacies that Africa will hand over to the world after the world cup tournament, since the Europeans, Americans and Asians could not resist the temptation of using it and are seen holding it to watch their matches. {3} Have you ever found yourself in bed in a dark room with just a single mosquito for company? The buzzing sound of the vibrations made by the mosquito's wings. {10} On the other hand, its ban will affect the mood of the host nation and, of course, other African countries at the world cup, because of the deep rooted emotions attached to it by fans. {3} This has sparked another controversy in the course of the tournament and has become the single item for discussion in the media since the LOC made that controversial statement on Sunday evening. {3}

Example 2 | swine flu india (2) | n: 25 | cohesion: 4.52 (Rank 4)
Patnaik, who created the image with the help of his students, on the golden beach has depicted the pig wearing a mask with the message 'Beware of swine flu'. The sculpture was put on display late Thursday on the beach in Puri, 56 km from the state capital Bhubaneswar. {18} Of the six cases reported in Pune, three are students who contracted the virus in the school. {91}

Example 3 | the lost symbol (4) | n: 20 | cohesion: 40.02 (MRV)
I read the book as fast as I could. Of course as a Dan Brown classic, it was very interesting, exciting and made me wanting to read as fast as I could. {13} Every symbol, every ritual, every society, all of it, even the corridors and tunnels below Washington, DC, it's all real. {3} I feel more connected to the message of this book (the reach and the power of the human mind) than I did to possibility that Jesus had a child. {12} Malakh is after secret knowledge guarded by the Masons and he'll stop at nothing to get it. To that end he's kidnapped Peter Solomon, the head of the Masonic order in Washington, DC. {1} Malakh is about to reach the highest level in the Masonic order, but even so, he knows he will not be admitted to the most secret secrets. Secrets that he's sworn to protect. He is not what he seems to his fellow Masons. He's lied to them. He has his own agenda. {2} [sic]

The first and third examples were ranked first (MRV) by our framework and the second one was ranked fourth. If we

examine the first example, a user who does not know the term 'vuvuzela' can immediately glean that it is a plastic trumpet which caused quite a stir in the fifa world cup 2010. There are also some sentences which insinuate toward a likely ban and surrounding controversy. In an ideal scenario, we would like to group sentences about the ban and the controversy in another view, but as it stands now, our view describes the instrument and the impact of vuvuzela on the world cup and serves as a good introduction to a novice or as a concise issue capsule to a user who is already familiar with the topic.

Similarly, the second example which was ranked fourth by our framework talks about the repercussions of the disease swine flu on pune and puri (cities in India). The third example, ranked first, contains some positive opinions about the book 'The Lost Symbol' and also a sneak peek into the intentions of the character Malakh. *Additional example views are provided in the appendix.*

The average number of sentences across all the views was found to be 3.9 and the average number of views across all the datasets was found to be 4.88. Table 4 presents the breakup for each dataset. Mean (S) indicates the average number of sentences across all the views, and Mean (N) indicates the average number of views. The implementation of the framework as described in Algorithm 1 took an upper-bound of 4.2 seconds to run, with computeTFIDF and calculateImportance being the time consuming steps at 2.6 seconds.

The main difference between summarization and our framework is that we provide multiple diverse views as opposed to summarization which lacks such an organization. We also rank these views thereby allowing a user to just look at the Most Relevant View (MRV) or the top x views as per his convienience. As we provide the IDs of the source articles in each view, a user can also browse through them to know more about that view.

4. CONCLUSION

Users who want to browse the content of a topic on the world wide web (www) have to wade through diverse articles available on it. Though summarization is successful in condensing huge volume of information, it groups several issues pertaining to a topic together and lacks an organized representation of the underlying issues representing it. In this paper, we propose a framework to mine the multiple views addressed by a collection of articles. These views are easily navigable and provide the user a detailed snapshot of their topic of interest. Our framework extends the concept of clustering to the sentence or phrase level (as opposed to document clustering) and groups semantically related sentences together to organize content in a way that is different from text summarization.

In future, we want to determine the polarity of a view (positive/negative/neutral) by examining the adjectives in it. We also want to incorporate user feedback by means of clicks, time spent on a page (implicit) and ratings, numerical scores (explicit) to evaluate the performance of our framework and if possible, re-rank the views.

Dataset	:	Number of TRS
financial meltdown	:	25
swine flu india	:	25
israel attacks gaza	:	30
the lost symbol	:	20
hotel taj krishna	:	20
hotel marriott	:	25
fifa vuvuzela	:	35
gulf oil spill	:	30

Table 3: Breakup of n

Dataset	Mean (S)	Mean (N)
financial meltdown	3.91	3.17
swine flu india	4.26	5.67
israel attacks gaza	3.74	4.17
the lost symbol	4.16	5.33
hotel taj krishna	3.67	4.17
hotel marriott	3.82	5.83
fifa vuvuzela	3.56	5.33
gulf oil spill	4.21	5.00

Table 4: Mean values

5. REFERENCES

[1] J. C. F. Allison L. Powell. Using multiple views of a document collection in information exploration. In *CHI'98: Information Exploration Workshop*, 1998.

[2] R. K. Ando, B. K. Boguraev, R. J. Byrd, and M. S. Neff. Multi-document summarization by visualizing topical content. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization - Volume 4*, NAACL-ANLP-AutoSum '00, pages 79–98, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[3] M. Charikar, C. Chekuri, T. Feder, and R. Motwani. Incremental clustering and dynamic information retrieval. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, STOC '97, pages 626–635, New York, NY, USA, 1997. ACM.

[4] Z. W. Department and Z. Wu. Verb semantics and lexical selection. In *In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, 1994.

[5] M. A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23:33–64, March 1997.

[6] R. Katragadda, P. Pingali, and V. Varma. Sentence position revisited: a robust light-weight update summarization 'baseline' algorithm. In *CLIAWS3 '09: Proceedings of the Third International Workshop on Cross Lingual Information Access*, pages 46–52, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[7] R. Mihalcea and C. Corley. Corpus-based and knowledge-based measures of text semantic similarity. In *In AAAI '06*, pages 775–780, 2006.

[8] A. L. Powell and J. C. French. The potential to improve retrieval effectiveness with multiple viewpoints. Technical report, VA, USA, 1998.

[9] M. Pucher. Performance evaluation of wordnet-based semantic relatedness measures for word prediction in coversational speech. In *IWCS 6:Sixth International Workshop on Computational Semantics Tilburg, Netherlands*, 2005.

[10] A. Tombros, J. M. Jose, and I. Ruthven. Clustering top-ranking sentences for information access. In *in Proceedings of the 7 th ECDL Conference*, pages 523–528, 2003.

APPENDIX

Example 4 | gulf oil spill (8) | n: 35 | cohesion: 16.64 (Rank 2) BP and the Coast Guard are also using chemicals to disperse the oil, which for the most part is spread in a thin sheen. But the area of the sheen has expanded to more than 150 miles long and about 30 miles wide. {1} The Coast Guard confirmed that the leading edge of the oil slick in the Gulf of Mexico is three miles from Pass-A-Loutre Wildlife Management Area, the Reuters news agency reported. The area is at the mouth of the Mississippi River. {1} "They're going to be focusing on the root cause, how the oil and gas were able to enter the [well] that should've been secured," he said. "That will be the primary focus, how the influx got in to the [well]." {1}

Example 5 | hotel marriott (6)| n: 30 | cohesion: 15.23 (Rank 3) Well located hotel offering good view of the lake. The rooms are clean and comfortable and have all amenities and facilities of a 5 star hotel. The hotel is not overtly luxurious but meets all expectations of a business traveller. The Indian restaurant is uper and a must-try. {6} The food is excellent and like I said, if it were not for the smell and so-so servie, I would stay here. {14} The rooms are great. Well lit, loaded with amenities and the trademark big glass windows to look out.. The bathroom is trendy and looks fabulous with rain shower and a bathtub. {12} [sic]

Example 6 | swine flu india (2) | n: 25 | cohesion: 15.98 (Rank 3) Three new cases of swine flu were confirmed in the city on Sunday, taking the total number of those infected to 12 in the State. {5} "Currently, it isn't the flu season in India, but if the cases keep coming in even after the rains, it will clash with our flu season (post-monsoon and winter period) which could be a problem", he said. {55} In Delhi, out of the four cases, three people, including two children aged 12, contracted the virus from a person who had the flu. {12}

Example 7 | financial meltdown (1) | n: 35 | cohesion: 0 (Outlier View) It has to be said: The model of the credit rating agencies has collapsed. Whether because of their unprofessionalism or inherent conflicts of interest, the fact that the agencies receive their pay from the companies they cover has bankrupted the system. {11}

Example 8 | israel attacks gaza (3) | n: 40 | cohesion: 0 (Outlier View) "I heard the explosions when I was standing in the hall for protection. Suddenly, in a few seconds, all of the police and firemen were in the building," said resident Rachel Mor, 25. {21}

ID	TRS	Mean (C)	Median (C)	V	O
1	20	17.58	17.6	3	10
	25	32.52	36.87	3	13
	30	11.23	11.23	2	15
	35	10.78	10.78	2	18
	40	12.5	16.74	3	20
	50	4.43	4.69	6	25
2	20	18.86	15.63	4	10
	25	15.6	15.6	4	13
	30	10.98	4.97	5	15
	35	14.16	5.12	7	18
	40	10.03	5.12	7	20
	50	11.42	4.79	7	25
3	20	13.32	4.52	3	12
	25	17.34	14.82	4	15
	30	19.38	21.56	4	18
	35	18.53	15.07	5	21
	40	7.11	5.1	4	24
	50	11.44	4.75	5	30
4	20	23.32	24.04	4	10
	25	16.55	16.3	6	13
	30	20.37	16.86	5	15
	35	7.18	5.07	5	18
	40	13.13	11.25	6	20
	50	8.46	4.54	6	25
5	20	10.61	10.61	2	10
	25	7.34	5.58	3	13
	30	5.48	5.58	3	15
	35	17.25	5.58	7	18
	40	12.02	6.59	4	20
	50	10.64	5.11	6	25
6	20	10.51	5.18	3	10
	25	19.83	15.23	5	13
	30	14.94	10.21	6	15
	35	14.55	10.37	6	18
	40	14	10.33	8	20
	50	7.87	4.47	7	25
7	20	11.52	5.09	4	10
	25	13.55	4.72	4	14
	30	11.9	4.73	5	16
	35	14.74	4.73	5	20
	40	8.35	4.59	6	26
	50	7	4.5	8	32
8	20	10.52	10.72	4	10
	25	13.95	10.55	4	14
	30	14.54	16.3	5	16
	35	12.94	10.61	6	19
	40	10.92	4.58	5	27
	50	11.81	4.72	6	34

Table 5: Results

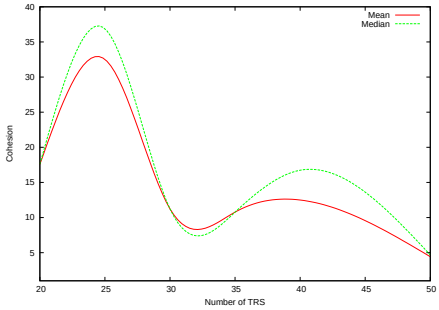

Figure 2: financial meltdown

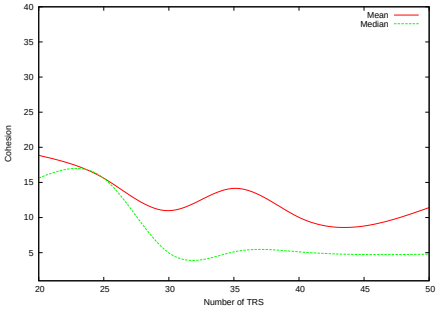

Figure 3: swine flu india

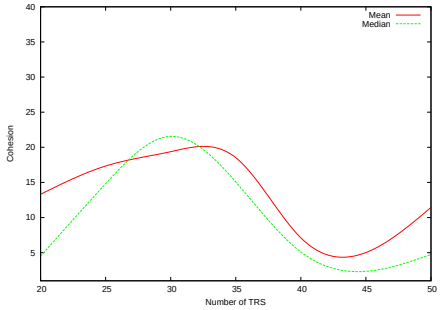

Figure 4: israel attacks gaza

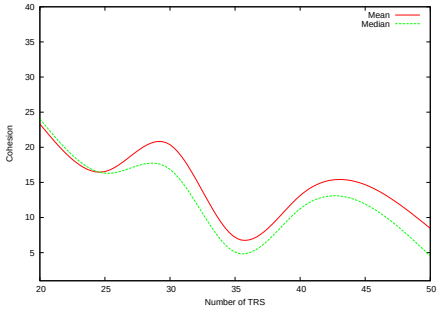

Figure 5: the lost symbol

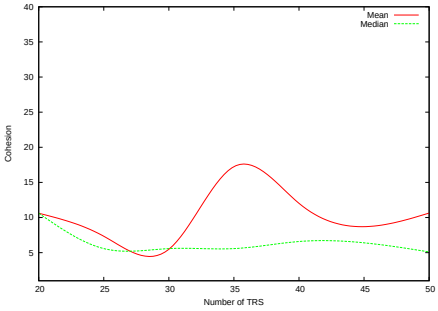

Figure 6: hotel taj krishna

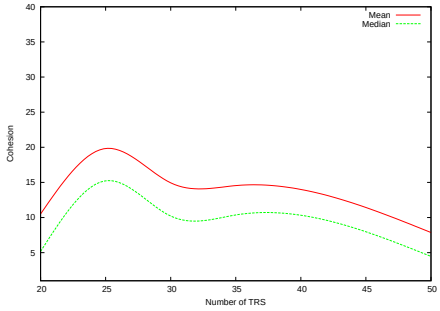

Figure 7: hotel marriott

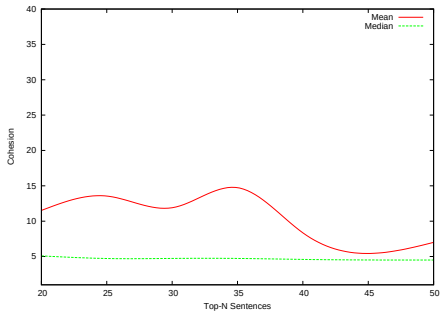

Figure 8: fifa vuvuzela

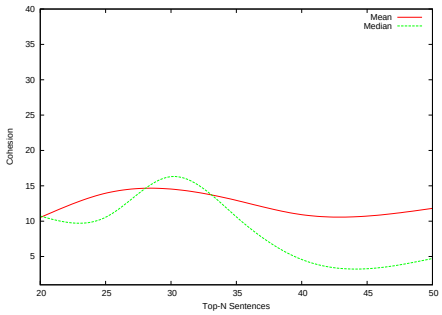

Figure 9: gulf oil spill