ISSN:0975-9646

A Novel approach for Novelty Detection of Web Documents

Manvi Breja

Computer Science & Technology, Manav Rachna University
Sector 43,Aravali Hills, Faridabad,Haryana

Abstract— — ***In order to reduce redundant and non-relevant information presented to users related to their query, there is a need for the novelty detection of those Web documents. This paper presents a novel approach to detect the novelty in the documents.***

Keywords— Novelty Detection, information retrieval, TREC, redundancy, information patterns.

I. Introduction

There is a continuous increase in the data volume that is uploaded and transmitted through the Internet between clients, services and Internet users [1]. People who work in media, security agencies receives a huge amount of stories, essays, reports and articles from a large number of sources. Such difficult situation inspired the researchers to invent new automatic system which is based on novelty detection. The novelty detection aims to build automatic systems which are capable to ignore old stories, essays, reports and articles already read or known, and notify the users of such systems about any new stories, essays, reports and articles.

An information retrieval systems (IRSs) is an information system, that is, a system used to store items of information that need to be processed, searched, retrieved and disseminated to various user populations [2]. Information retrieval is often referred as document retrieval. The basic task of document retrieval is to retrieve documents that are relevant to a user's request or information need. The output of a traditional document retrieval system is a ranked list of documents. Documents are ranked by the relevance scores that are calculated by the system. The system assumes that a document with a higher relevance score is more likely to be relevant to the user's request than a document with a lower relevance score. With the continuing growth of information, users of IRSs and search engines need to obtain useful information quickly; without the need to examine a lot of redundant information. Using the novelty detection the amount of redundant and irrelevant materials presented to users.

Novelty detection can be viewed as going a step further than traditional document retrieval. Based on the output of a document retrieval system (i.e., a ranked list of documents), a novelty detection system will further extract documents with new information from the ranked list. The purpose of the research on novelty detection is to provide a user with a list of text passages that both are relevant and contain new information with respect to the user's information need.

Novelty detection can be performed at three different levels: the event level, the document level and the sentence level. At the event level, a novel document is required to not only be relevant to a topic (i.e., a query) but also to discuss a new event. At the sentence level, a novel sentence should be relevant to a topic and provide new information. This means that the novel sentence may either discuss a new event or provide new information about an old event. Novelty detection at the sentence level is also the basis for novelty detection at the event level.

II. Related Work

A. Allan, Wade, and Bolivar [4] have presented a Topic Detection and Tracking (TDT) research and evaluation project; which is dedicated to novel online event detection and tracking. TDT tasks interested in inter-topic or inter-event novelty detection, in order to determine whether two news stories cover the same occasion. TDT is interested in with story-level online evaluation, where the news stories are presented one after another to be evaluated sequentially, and identify the new news stories.

B. Zhang, et al., have extended an adaptive information filtering system to make decisions about the novelty and redundancy of relevant documents [12]. They have proposed a set of five redundancy measures; with and without redundancy thresholds. The results of the conducted experiments proved that the cosine similarity measure and a redundancy measure based on a mixture of language models were effective techniques to identify redundant documents.

C. A number of studies exhibit how novelty detection can be used in several applications. A new formula and approach to the Minimal Document Set Retrieval (MDSR) problem was presented by Dai and Srihari [18], where three retrieval and ranking algorithms have been proposed and tested in their work. The three algorithms are: novelty based algorithm, cluster based method, and subtopic extraction based method.

D. Discovering new events automatically from chronologically ordered documents is a real challenge to researchers in this field. A particular form of novelty detection called First Story Detection (FSD); which is known as the most difficult in the field of Topic Detection and Tracking (TDT). FSD aims to detect on-line new news stories as soon as they arrive in the sequence of documents.

III. Introduction To Novelty Detection

Novelty Detection is the process of mining the novel but relevant information based on specific topic defined by user. It also called as novelty mining. It also defined as the opposite of redundancy i.e. given a set of relevant documents, any document which is very similar to history document is regarded as redundant.

A) Need of Novelty Detection

With the abundance of information, social networks, blogs, and news articles, many articles may contain similar information. Thus, there is an increasing need for identifying novel and relevant information out of a mass of incoming text documents, with past efforts made in text summarization, reuse detection, contextual information retrieval [19], text clustering . Novel information in this case refers to text that contain fresh content and novelty detection, is the process of singling out novel information from a given set of text documents [14]. Through this process, users can save time by reading only the novel information, while the repeated information is filtered away.

B) Levels Of Novelty Detection

Novelty detection (ND), or novelty mining, has been performed at three different levels:
event level [20], document level [8], and sentence level [21].

C) Novelty detection at the event level

Work on novelty detection at the event level arises from the Topic Detection and Tracking (TDT)[14] research, which is concerned with online new event detection and/or first story detection [3,7,8,9,10,11,5,13]. Several models, such as vector space models, language models, lexical chains, etc., have been proposed to represent incoming new stories/documents. Each story is then grouped into clusters. An incoming story will either be grouped into the closest cluster if the similarity score between them is above a preset similarity threshold or it will be used to start a new cluster. A story which starts a new cluster will be marked as the first story about a new topic, or it will be marked as ‘‘old’’ (about an old event) if there exists a novelty threshold (which is smaller than the similarity threshold) and the similarity score between the story and its closest cluster is greater than the novelty threshold.

D) Novelty detection at the sentence level

Research on novelty detection at the sentence level is related to the TREC novelty tracks [15, 16, 17]. The goal is to find relevant and novel sentences, given a query and an ordered list of relevant documents. Novelty detection at the sentence level can be conducted mainly in two steps: relevant sentence retrieval and novel sentence extraction. In current techniques developed for novelty detection at the sentence level, new words appearing in sentences usually contribute to the scores that are used to rank sentences. Usually a high similarity score between a sentence and a given query will increase the relevance rank of the sentence, whereas a high similarity score between the sentence and all previously seen sentences will decrease the novelty ranking of the sentence. The simplest novelty measure, New Word Count Measure [7], simply counts the number of new words appearing in a sentence and takes it as the novelty score for the sentence. There are other similar novelty or redundancy measures that consider new words appearing in sentences such as New Information Degree (NID) etc.

E) Novelty detection at the document level

Document-level novelty detection is more difficult than sentence level novelty detection. The reason is that nearly every document contains some new information, especially when the domain is new. However, a document is usually composed of several sentences so that the novelty of sentences in one document can help to judge the novelty of the document.

F) Intelligent Novelty Detection Techniques

Considering the diverse and changing scenario in the real world, there are many techniques for intelligent novelty mining for bridging the gap between the existing novelty mining methods and user performance requirements. Intelligent novelty mining addresses the domain-specific problem of mining novel information from text data with specific regard to the user context and aims to balance the technical significance and business concerns to create techniques that are useful in real-world scenarios. These techniques aim to adapt to the users desired level of novel information and human interaction. By addressing the issues of intelligent novelty mining, the techniques are useful from both the technical and business perspectives.

G) Level of Novelty

Different users have different definitions of novel information. For example, a user would regard a sentence with 50% novel information as a novel sentence while another user would regard a sentence with 80% novel information as a novel sentence. The threshold of novelty scores should be higher for the user with a stricter definition for the novel sentence. As novelty mining is an accumulating system, more training information will be available for threshold setting, based on the user’s feedback given over time.

IV. Comparision Of Our Approach With Others

In this paper, novelty detection at the document level is studied. A new approach for finding the novelty of the document is proposed. There are three main differences between our document based approach and the aforementioned approaches.

- First, traditionally, novelty detection always treats documents and sentences as two data sources and is performed separately. However, the prediction result of sentences can help to decide the novelty of the document. Before performing novelty detection, we need to first preprocess the documents by removing stop words, performing word stemming, etc.
- Second, by using this approach, a document which shares one single sentence with each of the history documents could be correctly recognized as redundant, whereas general document-level

methods may incorrectly label the document as novel.

- Third, this new framework can improve code reuse in a novelty detection system because we only need to focus on the development of the sentence-level modules.

V. Proposed Work

One of the most important points of concern is how the idea of novelty detection will refine existing search-engine results. Many important applications have used novelty detection in order to reduce redundant and non-relevant information presented to users of the document retrieval systems. In this paper, a novel approach to novelty detection for different web documents has been proposed.

- The proposed approach first connects with Google corresponding to give query given on user search engine and we retrieve the relevant results given by Google search engine.
- Then a Document to Sentence algorithm to novelty detection is applied.
- This algorithm is aimed at removing the redundancy of the results and increasing the speed to find the novel information in the documents.
- To increase the speed, novelty score of documents is calculated in the novelty detection algorithm.
- The novelty score is calculated by using the sentence segmentation instead of using whole document. Sentence segmentation has been done by preprocessing the document and then breaks the document into sentences.

VI. ARCHITECTURE OF PROPOSED NOVEL APPROACH

In the proposed system, user pass a query on his/her own search engine which is transferred to Google search engine which gives the relevant documents out of which information about initial 4 or 5 documents are fetched. On these documents, Novelty Detection algorithm is applied to determine novelty of a document.

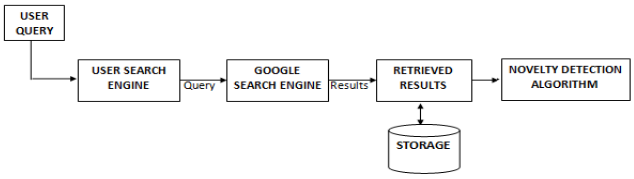


Fig.1.General Architecture of Proposed Novel Approach

A) User Search Engine

In user search engine, user can pass the query according to his/her requirement.

B) Google Search engine and Storage

User query then given to the Google search engine from where we extract information about first four pages and store in the Database (Storage).

C) Document level novelty detection algorithm

Document Level Novelty Detection (DND) algorithm is a proposed detection algorithm which is used to find whether a document gives novel information or not when compared with the history documents. DND first break the document into sentences, determines the novelty score of each document based on a fixed threshold.

For sentence segmentation a tool is used, Stanford parser, which break the document into sentences. Sentences are then compared with all the history sentences to compute the similarity between those sentences. To compute the nature of document, similarity is converted to novelty score for each sentence. A minimum value is selected out of the novelty values and finally the decision has to be made.

VII. Architecture Of Novelty Detection Module

In this system, user enters a query in the form of a new document; this is passed to segmentation module. This module breaks the documents into sentences and stored them in the database. Now the sentences are passed through text categorization module where stemming and stop word removal processing is done. Then the detector module decides the novelty of the document. At last the result is passed to the user by result module.

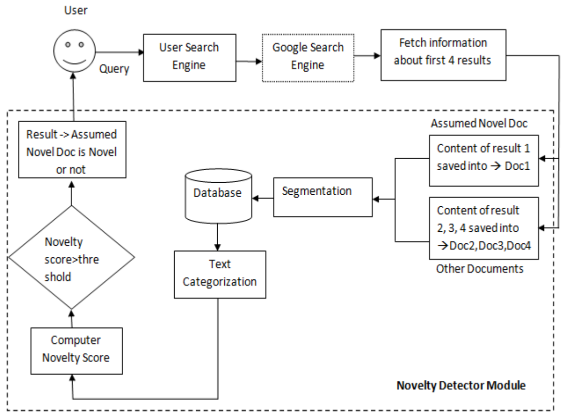


Fig.2. Architecture of Novelty Detection Module at Document level

Now, various components and novelty process are explained below in detail to have an understanding of the proposed detection system:-

A) Assumed novel document

User enter the query in the form of a new document, this document is compared with all the other (history) documents. Further processing is done by taking this document to the other modules.

B) Others Documents

This module contains the documents present in the database that are used for comparison with the new document.

C) Segmentation Module

This module breaks the documents into sentences and these sentences are stored in the database for further processing.

D) Text Categorization

This module helps in preprocessing the sentences. It removes the stop words from the sentences before they move to the next module.

E) Novelty Detector Module

This module helps in finding the novelty of the document. The document is segmented into sentences; compute the novelty score of each sentence by using the sentence-level novelty detection module. Then the average of novelty score is compared with a fixed threshold value, if the value of novelty score is greater than the threshold value then the document is considered as novel otherwise not.

There are several different geometric distance measures, such as Manhattan distance and cosine similarity. In this paper, cosine similarity is used to predict whether an incoming sentence contains enough novel information compared to a set of history sentences.

Cosine similarity is a symmetric measure related to the angle between two vectors. If we represent a sentence s as a vector s = [w1(s), w2(s), . . . , wn(s)]T , then the cosine similarity between two sentences is defined as:

$$\text{Cos (st, si)} = \frac{\sum wk(st)\ wk(si)}{||st||.||st||}$$

Where k=1 to n

Cosine Similarity measures the similarity between two sentences or documents in terms of the value within the range of [-1,1] whichever you want to measure.. Let me show you an example.

Let's consider two sentences:

1. Xeon goes to marry Xeonian girl, a girl.
2. Leon goes to forest to find Xeon.

From the first sentence, calculating the terms and their respective frequencies :

TABLE I
TERM AND FREQUENCIES OF SENTENCE 1 FOR CALCULATING COSINE SIMILARITY

TERMS	Xeon	goes	to	marry	Xeonian	girl	a
FREQUENCIES	1	1	1	1	1	2	1

If we do same for the second sentence,

TABLE II
TERM AND FREQUENCIES OF SENTENCE 2 FOR CALCULATING COSINE SIMILARITY

TERMS	Leon	goes	to	Forest	Find	Xeon
FREQUENCIES	1	1	2	1	1	1

In the above table the total number of terms in sentence 1 is 8 and in sentence 2 is 7.

Recall vector: Let's suppose : vector q = [2,2] and vector d = [0,1].

i.e. for above example it will be :

Now assuming you all knows what is cos product, now get what I am doing with terms in sentence 1 and sentence 2.

TABLE III
TERM AND FREQUENCIES OF SENTENCE 1 AND SENTENCE 2 FOR CALCULATING COSINE SIMILARITY

Term	Xeon	goes	to	marry	Xeonian	girl	Leon	Forest	Find
Freq in sentence 1	1	1	1	1	1	2	0	0	0
Freq in sentence 2	1	1	2	0	0	0	1	1	1

Then let : vec1 = [1,1,1,1,1,2,0,0,0] and vec2 = [1,1,2,0,0,0,1,1,1].

Therefore finally we get :

cos θ = 0.7071

Thus we can calculate the cos of angle between two vectors. Suppose vector q and d.

Then the **cos product** of vector q and d is :

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

$$\cos\theta = \frac{1\times 1+1\times 1+1\ \times 0+\cdots+0}{\sqrt{1^2+1^2+1^2+1^2+1^2+2^2}\ \times\ \sqrt{1^2+1^2+2^2+1^2+1^2+1^2}}$$

Where,

qi is the tf-idf weight of term i in the query

di is the tf-idf weight of term i in the document

cos(q,d) is the cosine similarity of q and d … or,

Equivalently, the cosine of the angle between q and d or Cosine similarity score between a and d

In the above formula, the tf-idf weight of a term is the product of its tf weight and its idf weight. It is the Best known weighting scheme in information retrieval.

The **log frequency weight** (tf weight) of term t in d is

$$w_{t,d} = \begin{cases} 1 + \log_{10} \text{tf}_{t,d}\ , & \text{if } \text{tf}_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases}$$

i.e. 0 → 0, 1 → 1, 2 → 1.3, 10 → 2, 1000 → 4, etc.

Score for a document-query pair: sum over terms t in both q and d is given as

$$\text{score} = \sum_{t\in q\cap d}(1+\log\ \text{tf}_{t,d})$$

The score is 0 if none of the query terms is present in the document.

We define the idf (inverse document frequency) [44, 45, 46] of term t by

$$\text{idf}_t = \log_{10}(N/\text{df}_t)$$

Thus tf-idf weight or score is calculated as

$$\text{w}_{t,d} = \log(1+\text{tf}_{t,d})\times\log_{10}(N/\text{df}_t)$$

It increases with the number of occurrences within a document and also increases with the rarity of the term in the collection.

For the novelty detection task, in order to measure the degree of novelty directly, we convert the cosine similarity score to the novelty score simply by (1-cosine similarity score). Cosine similarity metric compares the current sentence with each of its history sentences individually, where the minimum novelty score among them will be chosen as the novelty score of the current sentence.

Novelty Score (st) = min [1- cos (st, si)]

VIII. Algorithm For Document Level Novelty Detection

Algorithm 1: DND (N, Ass_novelDoc)

```
i/p: global set of N documents(Other documents): otherDoc, Assumed novel document : Ass_novelDoc
o/p Ass_novelDoc is novel or not
Begin
For i←1 to N
Begin
    For sen_doc←1 to sen_ Ass_novelDoc
    Begin
        For sen←1 to sen_idoc
        Begin
            Cos_sim[]←find cosineSimilarity(sen, sen_doc)
        End
        maxCos[]←find maximum value from cos_sim[]
        noveltyScore[]← min[1- maxCos[]]
    End
End
avgNovel ← ∑noveltyScore / N
If(avgNovel>Threshold)
Return Ass_novelDoc is novel
Else
Return Ass_novelDoc is not Novel
End
```

IX. Implementation And Result Analysis

The proposed system is implemented using various technologies like Java, PHP, HTML, Microsoft Access and CSS.

1. First start wampserver and open Search.php. Then query is given to the user search engine. The source code of this engine is written in PHP. "ABOUT" link provided in the bottom tell about the Novelty detection and approach used to determine novelty of a document. Suppose a query "Holi" is given.

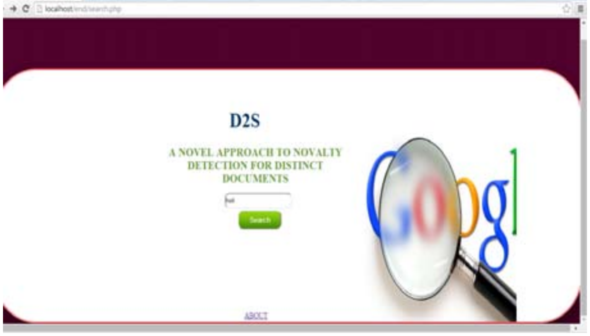
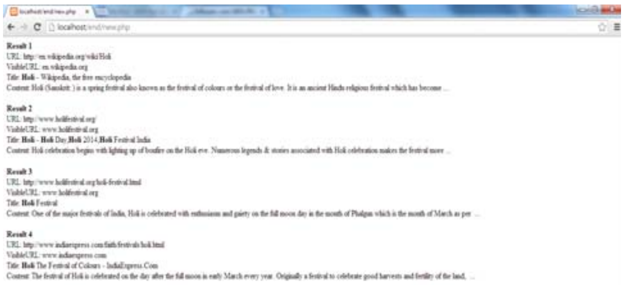

localhost/end/search.php
D2S
A NOVEL APPROACH TO NOVALTY DETECTION FOR DISTINCT DOCUMENTS
Search
ABOUT

Fig. 3. The User Search Engine

2. By clicking on "Search" button, it connect with Google search engine from where we extract information about first four webpage like URL, visible URL, title, content realted to query "Holi". Its code is also written in an PHP file.



Result 1
URL: http://en.wikipedia.org/wiki/Holi
VisibleURL: en.wikipedia.org
Title: **Holi** - Wikipedia, the free encyclopedia
Content: Holi (Sanskrit:) is a spring festival also known as the festival of colours or the festival of love. It is an ancient Hindu religious festival which has become ...

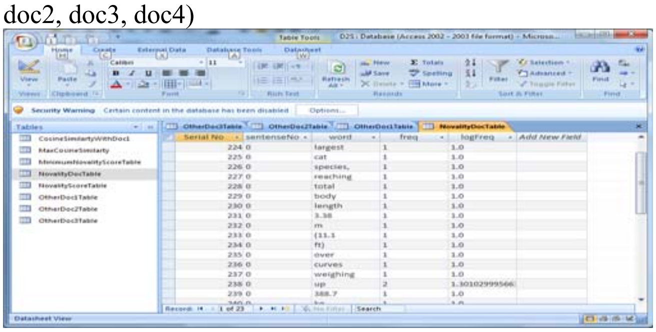
Result 2
URL: http://www.holifestival.org/
VisibleURL: www.holifestival.org
Title: **Holi** - **Holi** Day,**Holi** 2014,**Holi** Festival India
Content: Holi celebration begins with lighting up of bonfire on the Holi eve. Numerous legends & stories associated with Holi celebration makes the festival more ...

Result 3
URL: http://www.holifestival.org/holi-festival.html
VisibleURL: www.holifestival.org
Title: **Holi** Festival
Content: One of the major festivals of India, Holi is celebrated with enthusiasm and gaiety on the full moon day in the month of Phalgun which is the month of March as per ...

Result 4
URL: http://www.indianexpress.com/faith/festivals/holi.html
VisibleURL: www.indianexpress.com
Title: **Holi** The Festival of Colours - IndiaExpress.Com
Content: The festival of Holi is celebrated on the day after the full moon in early March every year. Originally a festival to celebrate good harvests and fertility of the land, ...

Fig 4. Extracting information about first four pages from Google retrieved results corresponding to a query

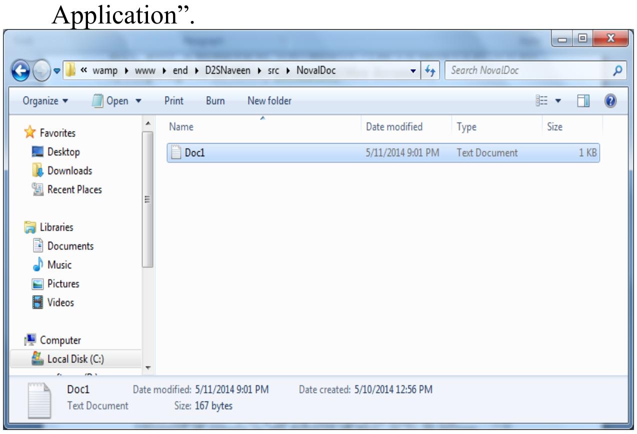
In step 2, at same time four documents are generated named doc1, doc2, doc3, doc4. "Doc1" is assumed as novel document i.e. we have to check it is novel or not by comparing with the History Documents(Other documents – doc2, doc3, doc4)



Serial No	sentenseNo	word	freq	logFreq	Add New Field
224	0	largest	1	1.0	
225	0	cat	1	1.0	
226	0	species,	1	1.0	
227	0	reaching	1	1.0	
228	0	total	1	1.0	
229	0	body	1	1.0	
230	0	length	1	1.0	
231	0	3.38	1	1.0	
232	0	m	1	1.0	
233	0	(11.1	1	1.0	
234	0	ft)	1	1.0	
235	0	over	1	1.0	
236	0	curves	1	1.0	
237	0	weighing	1	1.0	
238	0	up	2	1.30102999566	
239	0	388.7	1	1.0	

Fig. 5. Generated Document 1(Assumed Novel document)

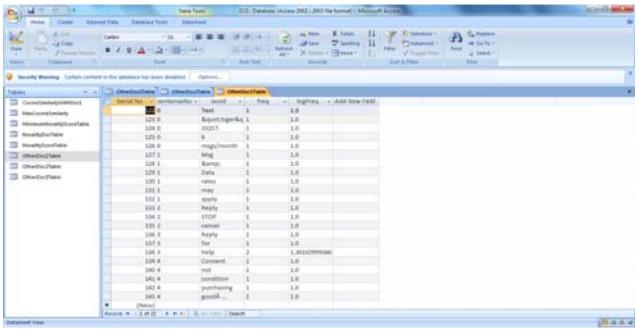
3. Open eclipse. Make a connectivity to ODBC and import the required project which is located in .../wamp/www/end And made in java. Here wamp stands form wampserver.
4. Run the project by clicking on "Run as Java Application".



Name	Date modified	Type	Size
Doc1	5/11/2014 9:01 PM	Text Document	1 KB

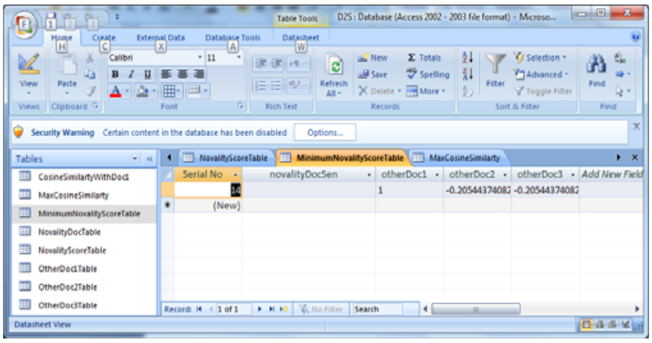
Doc1 Date modified: 5/11/2014 9:01 PM Date created: 5/10/2014 12:56 PM
Text Document Size: 167 bytes

Fig 6. Generated document1 (other document) table



Serial No	sentenseNo	word	freq	logFreq	Add New Field
[illegible]	0	Text	1	1.0	
123	0	&quot;tiger&q	1	1.0	
124	0	33257.	1	1.0	
125	0	6	1	1.0	
126	0	msgs/month	1	1.0	
127	1	Msg	1	1.0	
128	1	&amp;	1	1.0	
129	1	Data	1	1.0	
130	1	rates	1	1.0	
131	1	may	1	1.0	
132	1	apply	1	1.0	
133	2	Reply	1	1.0	
134	2	STOP	1	1.0	
135	2	cancel	1	1.0	
136	3	Reply	1	1.0	
137	3	for	1	1.0	
138	3	help	2	1.30102999566	
139	4	Consent	1	1.0	
140	4	not	1	1.0	
141	4	condition	1	1.0	
142	4	purchasing	1	1.0	
143	4	goodÂ ...	1	1.0	
(New)					

Fig. 7. Generated Assumed Novelty Document Table



Security Warning Certain content in the database has been disabled

Serial No	novalityDocSen	otherDoc1	otherDoc2	otherDoc3	Add New Field
14		1	-0.20544374082	-0.20544374082	
(New)					

Fig 8. Calculated Minimum Novelty Score Table

X. Conclusion And Future Work

A) Conclusion

The rapid growth in the amount of information and the number of users has lead to difficulty in providing effective search services for the web users. Due to the extremely large volume of documents on the web, analysis of the collection entire collection is not feasible. Novelty detection is an important activity used to identify new information, and reduce redundancy and the number of non-relevant information presented to users of systems; such as information retrieval systems, Web search engines, document filters and cross-document summarization. It can also be used by different tasks of natural language processing (NLP); such as machine translation, summarization, and question answering systems.

In this thesis, we proposed a new approach which felicitously applied document level novelty detection (DND) for document level novelty detection. This framework can make document-level novelty detection more effective by adopting the techniques for the sentence level. Our method DND can greatly improve the document level novelty detection performance in terms of redundancy-precision and redundancy-recall. Furthermore, DND seems better at finding redundant documents than novel information. These observations will be very helpful for successfully integrating DND to a real-world novelty detection system. Although this thesis presents the results for novelty detection, DND can be generalized for other types of document-level information retrieval.

B) Future Work

In future this study can be extended in following ways:-

- This architecture can be implemented to increase the efficiency of novelty detection method.
- Some other techniques, such as text summarization, can be tried in DND, which may possibly improve the novelty detection performance.
- Performance can be improved if we make use of various named entities and form multiple specific questions corresponding to query to obtain more relevant results.

References

[1] E. Greengrass, "*Information Retrieval: A Survey, DOD Technical Report TR-R52-008-001*", (2000).

[2] G. Salton and M. J. McGill 1983 *Introduction to modern information retrieval*. McGraw- Hill, ISBN 0070544840.

[3] J. Allan, R. Paka, and V. Lavrenko, "*On-line New Event Detection and Tracking*", Proc. SIGIR-98, 1998: 37-45.

[4] J. Allan, C. Wade, and A. Bolivar, "Retrieval and novelty detection at the sentence level," Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada, July 28-August 01, 2003.

[5] N. Stokes and J. Carthy, "*First Story Detection using a Composite Document Representation*", Proc. HLT01, 2001.

[6] J. Allan, R. Gupta, and V. Khandelwal, "*Temporal summaries of new topics*," Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States, p.10-18, September 2001.

[7] Y. Yang, J. Zhang, J. Carbonell and C. Jin, "*Topic-conditioned Novelty Detection*", SIGKDD, 2002: 688-693.

[8] Tsai FS, Chan KL (2010) *Redundancy and novelty mining in the business blogosphere*. Learn Organiz 17(6):490–499

[9] M. Franz, A. Ittycheriah, J. S. McCarley and T. Ward, "*First Story Detection, Combining Similarity and Novelty Based Approach*", Topic Detection and Tracking Workshop, 2001.

[10] J. Allan, V. Lavrenko and H. Jin, "*First Story Detection in TDT is Hard*", Proc. CIKM, 2000.

[11] Y. Yang, T. Pierce and J. Carbonell, "*A Study on Retro-spective and On-Line event detection*", Proc. SIGIR-98

[12] T. Brants, F. Chen and A. Farahat, "*A System for New Event Detection*", Proc. SIGIR-03,2003: 330-337.

[13] T. Brants, F. Chen, and A. Farahat, "*A System for New Event Detection*", in Proceedings of ACM SIGIR2003

[14] X. Li and W. B. Croft, "*Sentence level information patterns for novelty detection*", Ph.D. dissertation, University of Massachusetts Amherst, (2006).

[15] D. Harman, "*Overview of the TREC 2002 NoveltyTrack*", TREC 2002.

[16] I. Soboroff and D. Harman, "*Overview of the TREC 2003 Novelty Track*", TREC 2003.

[17] I. Soboroff, "*Overview of the TREC 2004 Novelty Track*", TREC 2004.

[18] W. Dai. and R. Srihari, "*Minimal Document Set Retrieval*," *Proc. ACM CIKM'05*, pp 752-759.

[19] Tamine-Lechani L, Boughanem M, Daoud M (2009) *Evaluation of contextual information retrieval effectiveness: overview of issues and research.*

[20] Obeid N, Rao RBKN (2009) *On integrating event definition and event detection*.

[21] Li X, Croft W B (2005) *Novelty detection based on sentence level patterns*. In: CIKM 2005 744–751