Comparison and Analysis of Two Approaches to Find Novel Documents out of Several Documents

Anjali Sharma
Department of Computer Science,
Meerut Institute of
engineering and Technology, UPTU

Mukesh Rawat, PhD
Assistant Professor,
Department of Computer Science,
Meerut Institute of
engineering and Technology, UPTU

ABSTRACT
Novelty detection system is used to extract documents with new or novel information from list of documents. Without looking for lot of redundant information, we can get useful information in a limited time. Cosine similarity and Language modeling are the two emerging techniques of information retrieval in today's scenario. The current study performs the analysis and comparison between these two models.

Keywords
Cosine similarity, Information retrieval, Language modeling, Novelty detection, Smoothing

1. INTRODUCTION
Novelty mining [1] or novelty detection [2] is the technique of finding relevant and novel information. As we know that there is volume of information available today, so it is difficult to find accurate information that suits user's need. When a user seeks for information on a particular topic, he looks at all the possible sources such as books, journals or articles and ends up with bulk of information. This problem of finding right information according to the need of user is solved using the process of information retrieval. For performing this purpose we use two techniques here cosine similarity [3] and language modeling [4].

2. BRIEF DESCRIPTION OF TWO MODELS

2.1.Cosine Similarity
Cosine similarity is the measure of similarity between two vectors or documents. This approach represents the vector space model [5] of information retrieval. This technique measures the cosine of the angle between two documents.

The cosine of two documents can be derived by using the Euclidean dot product formula:

$$A.B=|A|\,|B|\cos(\theta) \quad (1)$$

$$\cos(\theta) = \frac{A.B}{|A||B|} \quad (2)$$

cos (θ) is the similarity score calculated as follows

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (3)$$

Where, A_i is the tf-idf [6] weight of term *i* in the document A

B_i is the tf-idf weight of term *i* in the document B

Here cos(*A*,*B*) or cos(θ) is the cosine similarity of *A* and *B* or we can say Cosine similarity score between A and B. In the above formula, the tf-idf weight of a term is the product of its tf weight and its idf weight.

tf is the term frequency of term i in the document.

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

$$\text{tf wt} = 1 + \log_{10}(\text{tf}_{t,d}) \quad (4)$$

The similarity score is 0 if none of the terms of one document is present in other document.

We define the idf (inverse document frequency) of term t by

$$\text{idf}_t = \log_{10}(N/\text{df}_t) \quad (5)$$

N is the total number of documents in the collection

df_t is the document frequency of term t i.e. number of documents in which term t occurs.

(Note -A document is novel if its terms are also novel (previously unseen). This implies that the terms of a novel document have a generally high specificity and therefore high IDF values.)

Thus tf-idf weight is calculated as

$$w_{t,d} = \log(1+\text{tf}_{t,d}) \times \log_{10}(N/\text{df}_t) \quad (6)$$

After this similarity score i.e. cos(θ) is calculated.

And then finally Novelty score [2] of document A is calculated as

$$\text{Min}\,(1-\cos(A,B)) \quad (7)$$

If the novelty score of document is greater than the threshold [7] then document is said to be novel.

2.2.Language Modeling Approach
Language Modeling was first brought to information retrieval by Ponte and Croft (1998). Language model is a probability distribution over strings of text and represents the probabilistic model of information retrieval. As compared to cosine similarity approach, language modeling approach is based on finding probability rather than similarity for ranking of novel documents. Also term frequency, document frequency used in a different way in LM (Language Modeling).

Different techniques/models /approaches for finding probability distribution using LM are as follows:

1. Query likelihood model
2. Document likelihood model
3. KL- divergence model

In this paper we are going to use Query likelihood model for finding probability distribution.

2.2.1 Query likelihood scoring method:

Goal of this method is to determine which document best derives query. Documents that give a higher probability to the query indicate that they have more terms of query (term frequency).

Score (Q, D) in cosine similarity approach is defined as P (Q|D) in LM.

i.e. Score (Q, D) = P (Q|D)

Documents are then ranked based on their likelihood of generating that query. Here we use document in place of query i.e. we find probability distribution of terms of one document in other document (P (A|B)), where A and B are the two documents.

$$P(A|B) = \prod_{i=1}^{n} P(Ai|B) \quad (8)$$

$$P(A_i|B) = \frac{\text{tf Ai,B}}{|B|} \quad \text{MLE}$$

MLE (Maximum Likelihood Estimate) is defined as the term frequency (tf) of document A's terms appearing in document B divided by document B length.

But there are some problems associated with above formula: Whole similarity measure results in zero if any term of document A is missing from document B (this is called estimation problem) also Document B may be relevant to the document A but the document A's term is absent from the document B (this is called data sparseness problem).

To deal with estimation and data sparseness problem, smoothing [8] technique can be used. Smoothing smooth the probability estimates by lowering the probability estimate of the terms in document B and assigning probabilities to unseen terms in document B.

$$P(A|B) = \prod_{i=1}^{n}((1-\lambda)\frac{(tf_{Ai,Bi})}{|B|} + \lambda\frac{(tf_{Ai,C})}{|C|}) \quad (9)$$

$$\text{Smoothing} = \lambda\frac{(tf_{Ai,C})}{|C|}$$

λ is a parameter to control the amount of smoothing.
According to TREC [9] [10] (Text Retrieval Conferences) evaluations,
λ=0.1 for short queries
Term Frequency of documents:
λ=0.7 for long queries
Generally we take value of $\lambda = 0.5$

The U.S. National Institute of Standards and Technology (NIST) start this large Information Retrieval test evaluation series i.e. TREC since 1992.

$(tf_{Ai,Bi})$ is occurrence of terms of document A in document B

$(tf_{Ai,C})$ is occurrence of terms of document A in the collection

|C| = No. of terms in the entire collection

|B| = No. of terms in document B.

After calculating similarity measure i.e. P (A|B), novelty score is calculated as follows

Min (1- P (A|B))

Then finally this novelty score is matched with the threshold value. If this value is greater than this threshold value then document is said to be novel.

3. EXPERIMENTAL RESULTS

Let us take an example

Step by step procedure for finding novel documents among three documents using Cosine similarity approach -

Step1. First of all enter the query which is used to find the relevant documents. Suppose query is "Political Corruption"

Step2. Then the documents are as follows:

D1: "Political corruption adversely affects development of nation."

D2: "Political corruption is the use of powers by government officials for their personal profit."

D3: "A Corrupted politician affects life of innocent people by showing their powers."

Step3. Now divide these documents into separate words in order to find Term frequency, Inverse document frequency and Tf-idf weight in case of cosine similarity and collection frequency in Language modeling. Now these documents are divided into terms (after removing stop words) and term frequency is calculated in a following way:

Table 1.TF for D1

Terms	Political	corruption	adversely	affects	development	nation
tf	1	1	1	1	1	1
tf wt	1	1	1	1	1	1

Table 2.TF for D2

Terms	Political	corruption	use	powers	government	officials	personal	profit
tf	1	1	1	1	1	1	1	1
tf wt	1	1	1	1	1	1	1	1

Table 3.TF for D3

Terms	Corrupted	politician	affects	life	innocent	people	showing	powers
tf	1	1	1	1	1	1	1	1
tf wt	1	1	1	1	1	1	1	1

Step4.

Table 4.IDF weight for all the terms of collection

Terms	IDF
political	0.176
corruption	0.176
adversely	0.477
affects	0.176
development	0.477
nation	0.477
use	0.477
powers	0.176
government	0.477
officials	0.477
personal	0.477
profit	0.477
corrupted	0.477
politician	0.477
life	0.477
innocent	0.477
people	0.477
showing	0.477

Step5:

Table 5.TF-IDF weight ($w_{t,d}$)for document D1

Terms	Political	corruption	adversely	affects	development	nation
$w_{t,d}$	0.176	0.176	0.477	0.176	0.477	0.477

Table 6.TF-IDF weight ($w_{t,d}$) for document D2

Terms	Political	corruption	use	powers	government	officials	personal	profit
$w_{t,d}$	0.176	0.176	0.477	0.176	0.477	0.477	0.477	0.477

Table 7.TF-IDF weight ($w_{t,d}$) for document D3

Terms	Corrupted	politician	affects	life	innocent	people	showing	powers
$w_{t,d}$	0.477	0.477	0.176	0.477	0.477	0.477	0.477	0.176

Step6: Now calculate similarity score between two documents

Similar Matrix:

	D1	D2	D3
D1	1	0.282	0.199
D2	0.282	1	0.158
D3	0.199	0.158	1

Step7: Finally calculate novelty score of documents

Novelty score of documents is calculated using formula:

Min (1- cos(*A*,*B*))

Novelty Matrix:

	D1	D2	D3
D1	0	0.717	0.800
D2	0.717	0	0.841
D3	0.800	0.841	0

⇩

D1's novelty score =0.717 (minimum score)

Similarly, D2's novelty score = 0.717

D3's novelty score = 0.800

So, Document D3 is novel among all documents.

Let threshold = 0.75

So, only Document D3 is novel among all other documents according to threshold 0.75

Now calculate novelty score of documents using language modeling approach.

Step1: Calculate similarity score between two documents.

Take λ = 0.5 as discussed above

Similar Matrix:

	D1	D2	D3
D1	1	2.74E-4	9.15E-5
D2	5.06E-5	1	1.25E-4
D3	0.00425	0.0035	1

Step2: Calculate novelty score of documents using formula

Min (1- P (A|B)

Novelty Matrix:

	D1	D2	D3
D1	0	0.9997	0.9999
D2	0.9994	0	0.9998
D3	0.9957	0.9964	0

⇩

D1's novelty score= 0.9957 (minimum score)

Similarly, D2's novelty score =0.9964

D3's novelty score =0.9998

So, document D3 is novel among all documents.

When we check novelty of documents according to threshold value 0.75 in case of LM, we came to know that all the three documents are novel. This is clear in the graphs given below:

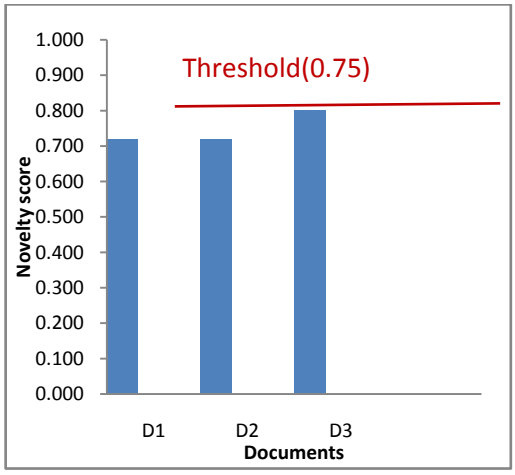

Fig.1 Graphical representation of document's novelty score in CS approach

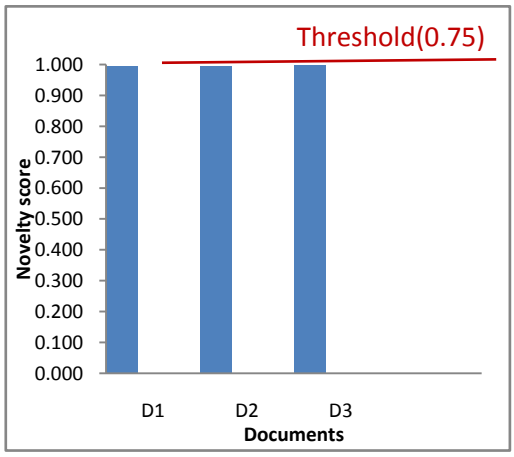

Fig.2 Graphical representation of document's novelty score in LM approach

4. COMPARISON BETWEEN THE TWO MODELS

Table 8.Comparison between CS and LM

Cosine Similarity Approach	Language Modeling Approach
• CS (Cosine Similarity) was first proposed by Salton & McGill in 1983.	• LM (Language Modeling) was first brought to information retrieval by Ponte & Croft in 1998.
• It uses geometrical tools to model the documents and terms.	• It is a branch of probabilistic models.
• IDF (Inverse Document Frequency) technique is used for the purpose of smoothing.	• Collection frequency is used for smoothing.
• A document is viewed as a vector and terms as elements of vector.	• A document is viewed as a language model.
• Numbers of computations are much more than LM approach.	• Lesser number of computations is done.
• Novelty score of documents is generally less than LM.	• Due to high novelty score, number of novel documents is also more than CS approach.

5. CONCLUSION

Cosine similarity and Language modeling approach both are the novelty measurement techniques for finding novel documents out of list of documents, but both uses different mathematical tools for this purpose. In this paper we came to know that in language modeling approach novelty score of documents is generally higher than the novelty score in cosine similarity approach. So we get higher number of novel documents in Language modeling approach. Also Cosine similarity approach is computationally more complex than Language modeling approach. Therefore, we conclude Language modeling approach better than Cosine similarity approach for finding novelty of documents. For future work, we will focus improving performance of Language modeling technique in various fields like text summarization, user modeling web search, classification, term distribution. After focusing on these areas language models will yield better results.

6. REFERENCES

[1] Flora S. Tsai, Review of techniques for intelligent novelty mining, Information Technology Journal (6): 1255-1261, 2010

[2] Manvi Breja, A novel approach for novelty detection of web documents, International Journal of Computer Science and Information Technologies, Vol. 6 (5), 2015, 4257-4262

[3] Ming-Feng Tsai, Ming-Hung Hsu, and Hsin-Hsi Chen, Similarity computation in novelty detection, Department of Computer Science and Information Engineering National Taiwan University, Taiwan.

[4] Djoerd Hiemstra, Language Models, In M. Tamer Özsu and Ling Liu (eds.) Encyclopedia of Database Systems, Springer, ISBN 978-0-387-49616-0, pages 1591-1594, 2009

[5] Jitendra Nath Singh and Sanjay Kumar Dwivedi, A Comparative Study on Approaches of Vector Space Model in Information Retrieval, International Journal of Computer Applications (0975 – 8887) International Conference of Reliability, Infocom Technologies and Optimization, 2013

[6] Stephen Robertson Microsoft Research 7 JJ Thomson Avenue Cambridge CB3 0FB UK Understanding Inverse Document Frequency:On theoretical arguments for IDF, 2004

[7] Xuchang Zou, Raffaella Settimi, Jane Cleland-Huang, Chuan Duan, Thresholding Strategy in Requirements Trace Retrieval

[8] CHENGXIANG ZHAI and JOHN LAFFERTY, Carnegie Mellon University, A Study of Smoothing Methods for Language Models Applied to Information Retrieval, ACM Transactions on Information Systems (TOIS), Volume 22 Issue 2, April 2004 Pages 179-214

[9] Flora S. Tsai · Yi Zhang, Document-to-sentence framework for novelty Detection, Knowl Inf Syst Volume 29 Issue 2, November 2011 Pages419-433

[10] Ian Soboro and Donna Harman National Institute of Standards and Technology Gaithersburg, MD, Novelty Detection: The TREC Experience, HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing Pages 105-112