Using Cohesion-Model for Story Link Detection System

K.Lakshmi, Saswati Mukherjee

Anna University, Chennai, India.

Summary

Cohesion-Model is a new Story Link Detection (SLD) System that is inspired by the relevance model of TDT. Task in hand is to detect whether given two documents are linked. Each document is expanded using corresponding relevant documents. Each term in the relevant document is weighted according to the cohesion factor. The two models then are compared using the modified fractional similarity. Performance of this model shows distinct improvement when compared with most effective Link Detection System, which measures similarity between the given two documents using cosine method. The model is also compared with a system using modified fractional method without building Cohesion-Model. The experimental results show the performance of Cohesion-Model tested with TDT4 data and proves the effectiveness of the new model.

Key words:

Information Retrieval, Information Organization, Story Link Detection, Topic Detection and Tracking, Similarity Measure.

1. Introduction

Topic detection and tracking (TDT) is a research area concerned with organizing a multilingual stream of news broadcasts as it arrives over time. In TDT an event is defined as something that happens at some specific time and place [3]. For example, a story about a tornado in Kansas in May and another story about a tornado in Nebraska in June should not be classified as linked because they are about different events, although they both fall under the same general topic of natural disasters. But a story about damage due to a tornado in Kansas and a story about the clean up and repairs due to the same tornado in Kansas are considered linked events [4]. In TDT a topic is defined to be a set of news stories that are strongly related by some seminal real-world event [1,2].

TDT investigations include five different tasks:

Story Segmentation: Segmenting a stream of data into distinct stories
First Story Detection: Identifying those news stories that are the first to discuss a new event occurring in the news
Cluster Detection: Given a small number of sample news stories about an event finding all following stories in the stream.
Tracking: Monitoring the news stream for finding additional of new stories to the existing topics.
Link Detection: is to decide whether the two randomly selected stories discusses about the same topic.

Most TDT tasks have at their core a comparison of two text models. In *story link detection*, comparison is between pairs of stories, to decide whether given pairs of stories are on the same topic or not. In *topic tracking*, the comparison is between a story and a topic, which is often represented as a centroid. Model that is developed for link detection itself can be used for topic tracking by changing the document model to a topic model. Given an event, tracking all the news stories that are relevant to that event helps the user to track the event of his interest. For detecting *new event*, link between the new story and the existing news stories can be identified. New story is compared with the existing stories to find whether they are linked. When there is no link between existing stories, new story is identified as new event (first story detection). Thus Link Detection system is considered as a core component for all other TDT tasks [5, 6].

1.1 Issues

TDT is different from usual query based retrieval. In Information Retrieval, users are interested in only top *n* URLs. In TDT, users intend to read all news articles relevant to a particular event. Also TDT systems are supposed to operate on live news feeds; decision on the incoming document has to be taken immediately (i.e. whether they are linked). This imparts additional challenges of algorithmic design.

In application such as text classification, summarization, we deal with a static collection of documents. The main advantage of a static collection is that we can learn a statistical model once, and then use it for all queries issued against that collection. In TDT, the models have to grow and adapt to match the ever-changing nature of a live news-feed. The users need to be alerted to events of interest as soon as possible, so for every document in the stream we need to make a hard decision of whether this document is relevant or not.

Another challenge in TDT is that the news article comes from different sources and languages [6, 7]. The streamed stories that originate in different languages are also available in English translation. The translations would

Manuscript received March 5, 2007
Manuscript revised March 25, 2007

have been performed automatically by machine translation algorithms, and, as a result, are inferior to manual translations. Similarly broadcast news is automatically converted to text form. While converting speech to text documents, many terms, especially the names are misspelled. This makes the comparison of document pairs more difficult.

1.2 Proposed Model

Reference [12] has studied the usage of communication and collective action for the development of society. One of the important parameter in this concept is the social cohesion of the members in the society. Social cohesion, explores how member of the society and the interconnection between them help in achieving any predefined goal. Social cohesion includes six characteristics that include sense of belonging, feelings of morale, goal consensus, trust, reciprocity, and network cohesion.

We find a strong connection between the social cohesion and cohesion among the terms in the documents. We propose a model that is inspired by the social cohesion concept. Query expansion is a concept used in IR and in Link Detection System. Our proposed model combines query expansion technique with the concept of social cohesion.

With this background information we have organized this paper with section 2 discussing about the related work. Section 3 elaborates the proposed system. Section 4 is about the experiment results. Next section concludes the results and discusses further possible enhancements.

2. Related Work

A number of research groups have developed story link detection systems. The best current technology for link detection relies on the use of cosine similarity between document terms vectors with TF-IDF term weighting. Reference [11] has used the concept of query expansion, a well established technique in IR. Here document given for link detection is considered as query and the documents that are relevant are retrieved from the collection. By using local context analysis technique terms in the relevant documents are added to the given document (query). Thus each document is expanded and then compared. This technique showed slight improvement over the link detection system that does not use any query expansion. However success of the model depends on how the relevant documents are fetched and how the terms are weighted.

We are strongly influenced by the success of query expansion technique in link detection. However the basic issues in query expansion used in link detection are i) How to expand the query document ii) How to assign weight to the terms iii) comparison of expanded models.

Reference [5,6] has used probability method for obtaining the relevant documents. Given a query document Q, probability of retrieving document (in the collection) D i.e., P(D|Q) is calculated. Top n documents that have high probability score are considered as relevant documents. Each term in the relevant documents are given weight according to the term's probability in the document and the probability of the document to be retrieved for the given query. These two topic models were compared by using Clarity adjusted Kullback-Leibler divergence method.

Another novel technique was used by [8] for link detection system. Work proposed by [8] is to use weighted output of a set of similarity metrics like cosine, Hellinger, Tanimoto, and clarity. In addition they have also proposed to use source-pair information in the link detection system and proved the improvement in the performance both the cases.

In the language models used term are considered to be independent of each other, which is not true in the real case. Reference [9] intends to use term dependency to capture the underlying semantics in the document. They proposed modeling sentences, rather than words or phrases as individual entities.

Using "soundex" in comparing documents of different sources (broadcast and newswire) is proposed by [13]. When broadcast news is converted to text, most of the nouns are given different spelling. So when a broadcast news and newswire news are compared some of the terms doesn't match and leads to low similarity value. System proposed by [13] tried to address this problem, but not able to produce better result due to poor named-entity recognizer for ASR documents.

3. Cohesion-Model

Documents given for link detection system are first expanded with the terms in the relevant documents. Each term is then assigned a weight according to its relevance and cohesiveness. The expanded models are compared to decide whether the given two documents are linked.

For comparing the documents we have used Modified Fractional Similarity measure proposed by [10]. Modified Fractional Similarity measure gives credit for the overlapping term and reduces the similarity score for having non-overlapping terms. So the similarity score depends on both similar as well as non-similar terms. With the encouraging performance in text classification, we use Modified Fractional Similarity as given in equation (1) for comparing the query document and the document in the collection. Modified Fractional Similarity provides a better precision and reduces false positives in the retrieved document. This helps to build improved link detection as shown by experimental results.

To achieve link detection using the proposed method, the first step is to obtain the documents relevant to the given documents. This retrieval is achieved by considering the given documents as query. This process is explained in the following sub-section.

3.1. Relevance Set

Documents that are relevant to the query document D1 are retrieved from the collection. Only documents that appear before D1 are considered. Searching the entire document in the collection could be time consuming. So to reduce the processing time we limit the search to *n* days. This look-ahead period is called as deferral period. Also it is sufficient to look the documents that are temporally closer to the query document as the events are time specific. Here deferral period is taken as 15. After retrieving document within the deferral period, each document is compared with the given (query) document. Documents that are greater than the threshold are considered to be relevant documents. Empirically we have set threshold as 0.15. Equation (1) is used to find similarity between the query document and all the documents within the deferral period. At most top 10 similar documents are considered as relevant documents.

$$\left.\begin{array}{ll} Mfraction(d1,d2) = 2*\alpha/(\beta+\gamma) & \\ \qquad\qquad if\ \{d1\} - \{d2\} \neq \varphi & \\ \qquad = \alpha & \\ \qquad\qquad if\ \{d1\} - \{d2\} = \varphi & \\ \alpha = \sum_{i=1}^{p} w_i * v_i & if\ term_i \in d1\ and \in d2 \\ \beta = \sum_{i=1}^{m} w_i & if\ term_i \in d1\ and \notin d2 \\ \gamma = \sum_{i=1}^{n} v_i & if\ term_i \in d2\ and \notin d1 \end{array}\right\} \quad (1)$$

w_i – weight of term$_i$ in d1,v_i – weight of term$_i$ in document d2, m – number of terms in the d1,n – number of terms in the document d2, and p – number of terms in both d1 and d2

We agree with [11] that cost of expanding using non-relevant passages is very high; the query will be expanded in a direction that is not related to the original request. The relevant document set includes some false positives also. Including the terms present in the false positive documents affects the performance of the system. To avoid this problem we intend to build Cohesion-Model based on the term's cohesiveness towards given document and also with the other terms in the relevant document set.

3.2. Building Cohesion-Model

Reference [12] describes an iterative process where "community dialogue" and "collective action" work together to produce social change in a community that improves the health and welfare of all of its members. It discusses Social Cohesion to be one of the factors affecting the Social Change. Social cohesion is an important antecedent and consequence of successful collective action.

Social cohesion consists of the forces that act on members of a group or community to remain in, and actively contribute to, the group. This idea of social cohesion fits well with the term cohesion needed for the model representing the given document. We map the social cohesion to the term cohesion in the relevant documents. Thus the model created will reflect the cohesion among the terms in the relevant documents.

Of the six factors mentioned in social cohesion, we find four to strongly influence the process of document expansion as applied to Link Detection System.

- Sense of belonging
- Feelings of morale,
- Goal consensus,
- Network cohesion.

Sense of belonging - is the extent to which individual members feel as if they are an important part of the group or community. This can be directly mapped with the term's frequency *tf* in the relevant documents. It is calculated using equation 2.

$$tf(t) = [1/N] * \sum_{d \in rl} tf(t,d) \qquad (2)$$

Where
tf (t) = term frequency of terms in relevant documents, rl – relevant documents, tf(t,d) = term frequency of term t in document d, and N – Total no of relevant documents

Feelings of morale – This refers to the extent to which members of a group or community are happy and proud of being a member. We can map this to the *inverse document frequency* (*idf)* factor of the term. Presence of a term in all or most of the documents across the boundaries of groups in the collection can be viewed as lack of confidence and lack of enthusiasm to identify itself with the group.

Goal consensus – It is the degree to which members of the community agree on the objectives to be achieved by the group. Here we translate it as how many times each term is repeated in relevant documents, which is denoted as document frequency. We calculate the *document frequency* (df) of the term in the relevant document set using equation 3. More number of times the term is repeated more overlap is expected among the relevant document set.

$$df(t) = [1/N] * docfreq(t) \qquad (3)$$

Where
df (t) =document frequency of terms t, docfreq –no.of times term t appears in relevant documents, N – Total no of relevant documents

Network cohesion - This can be viewed as the term's co-occurrence in the relevant documents. By adding co-occurrence weight in calculating the terms weight is expected to eliminate the problem described in [11]. Even though the quality of the retrieval is poor i.e., the retrieved document set contains negative documents, terms of the negative documents may not co-occur with the positive document terms. Equation 4 is used to calculated the co-occurrence weight

$$Cnet(t_i) = \sum_{ti,tj \in \{T\}} n(t_i \cap t_j)/n(t_i) \qquad (4)$$

Where
T – All terms in the relevant documents,Cnet (t_i) - Cohesion value of term t_i, $n(t_i \cap t_j)$ – no of times terms t_i & t_j co-occurred in relevant documents, and $n(t_i)$ = document frequency of term t_i in relevant documents

With all the parameter final weight of the term is calculated as given in equation 5.

$$w(t_i)=c1*tf(t_i)+c2*idf(t_i)+c3*df(t_i)+c4*Cnet(t_i) \qquad (5)$$

Where
w (t_i) - weight of term t_i,T – All terms in the relevant documents, tf (t_i) = term frequency of terms in relevant documents, idf (t_i) = term frequency of terms in relevant documents, d(t_i) = document frequency of term t_i in relevant documents, Cnet (t_i) - Cohesion value of term t_i, and c1-c4 - constants

Each term is given weight as the weighted sum of the tf, idf, df and Cnet. Equation 5 shows how the term weight of each term is calculated. Constants c1-c4 are selected empirically.

4. Experimental Results

In this section we evaluate performance of Cohesion model, as described above, on the Link detection task of TDT. First, we describe the experimental setup and the evaluation methodology

4.1. Experimental Set up

We have used TDT4 data for evaluating our proposed system. We have considered 16 topic's data for the experiment. Test data contains 377 positive links and 1277 negative links. The news stories were collected from different sources newswire sources (Associated Press and New York Times) and broadcast sources (Voice of America and Public Radio International). We consider only English stories for our experiments. Stories of other languages are not considered for these experiments. Text version of the broadcast news is included in this evaluation. Data set used for experimentation includes documents belonging to different news sources (Newswire and Broadcast).

4.2. Evaluation Method

The system is evaluated in terms of its ability to detect the pairs of stories that discuss the same topic. During evaluation the Link Detection System emits a YES or NO decision for each story pair. If our system emits a YES for an off-target pair, we get a False Alarm error; if the system emits a NO for on-target pair, we get a Miss error. Otherwise the system is correct.

Link Detection is evaluated in terms of F1-Measure as in classification system or Cost Function, which is a weighted sum of probabilities of getting a Miss and False Alarm [9]. Cost is calculated by using equation 6.

$$Cost = P(Miss)CMiss + P(FA)CFA \qquad (6)$$

Here we have considered F1- measure as the main factor for evaluating the performance of the systems. We have chosen F1- measure because we want to use the system as a basic component of TDT as mentioned earlier. In [8] Chen et al., has shown that optimized story link detection is not equivalent to optimized new event detection. An optimal link detection system tries to reduce the false alarm (as the weight of the false alarm is high). But false alarm of Link Detection system is equivalent to miss in New Event Detection System (NED). Thus an optimized Link Detection System does lead to optimized NED. So we have used F1- measure to indicate the performance of the Link Detection System, as F1- measure is a harmonic mean of precision and recall.

An operational Link Detection System requires a threshold selection strategy for making YES / NO decisions. However, in a research setting it has been a common practice to ignore on-line threshold selection and perform evaluations at the threshold that gives the best possible cost. Here we have considered the break even F1-measure i.e., F1- measure calculated when both precision and recall are equal. Cost and Accuracy are also calculated at the break-even point.

4.3. Experimental Method

Only preprocessing done here is stopword removal. Stemming is not used in preprocessing the documents. We have taken Link Detection System that uses cosine similarity to compare the query documents as the base method to compare our proposed models. We have proposed methods with and without query expansion technique.

Link Detection Systems we have considered can be broadly classified as system with query expansion technique and without expansion. As given in table 1, System 1 and 2 are without query expansion and all the others are involved with query expansion technique.

Table1. Various Link Detection Systems used for experiments

Without Query Expansion	1. cs - Cosine Similarity Method
	2. mf - Modified Fractional Similarity Method
With Query Expansion	3. tf
	4. tfidf – tf*idf
	5. tfdf – tf*df
	6. tfidfdf – tf*idf *df
	7. ltfdf –linear tf df
	8. ltfcnet – linear tf cnet
	9. ltfidfdf – linear tf idf df
	10. ltfidfcnet – linear tf idf cnet
	11. ltfdfcnet – linear tf df cnet
	12. ltfidfdfcnet – linear tf idf df cnet

The cosine similarity given by equation 7, is a classic measure used in Information Retrieval, and is consistent with a vector-space representation of stories. The measure is simply an inner product of two vectors, where each vector is normalized to unit length. It represents the cosine of the angle between the two vectors d1 and d2. Cosine similarity tends to perform best at full dimensionality, as in the case of comparing two long stories. Performance degrades as one of the vectors becomes shorter. Because of the built-in length normalization, cosine similarity is less dependent on specific term weighting, and performs well when raw word counts are presented as weights.

$$\text{Cos}(d1,d2) = \Sigma\, d1.d2/|d1||d2| \qquad (7)$$

Modified similarity measure has proposed by [10] indicates a better performance over cosine similarity. So we inclined to use modified similarity measure in link detection system. Similarity between the given two documents is calculated by using equation (1) in section 3.1.

Systems 3-12 are constructed with various factor of social cohesion concept. First we take the basic parameter *tf* and then combining other factors one by one to evaluate the contribution of each factor for over all system performance. Thus System 3 consists of only *tf* factor, which contributes for sense of belonging. System 4 is a combination of *tf* and *idf*, where *tf* contributes sense of belonging and *idf* for sense of moral. System 4 is constructed with *tf***idf*. System 5 is a combination of *tf* and *df*, where tf is for sense of belonging and *df* is given for goal consensus. System 5 is given as *tf***df*. System 6 is constructed with *tf***idf* **df*, where *tf* is for sense of belonging, *idf* for sense of moral and *df* for goal consensus. Systems 3-6 are constructed with generative effect of different factor whereas Systems 7-12 are constructed with linear combination effect of various factors. System 7 is a linear combination of tf and df constructed using c1*tf+c2*df.

System 8 is a linear weighted sum of tf and cnet constructed as c1*tf+c2*cnet. Weight c1 and c2 are constants, assigned values empirically. System 9 is a linear combination of tf, idf and df. It is constructed using c1*tf+c2*idf+c3*df. System 10 is a linear combination of tf and idf and cnet. It is constructed with c1*tf+c2*idf+c3*cnet. System 11 is a linear combination of tf, df and cnet. It is constructed with c1*tf+c2*df+c3*cnet. System 12 is a linear combination of tf, idf, df and cnet. It is constructed with c1*tf+c2*df+c3*cnet

Table 2 shows the break even F1- measure and corresponding cost and accuracy for the various link detection systems. Figures 1-3 show comparison of f1 measure, accuracy, and cost of various link detection systems respectively.

Table 2. Performance of Link Detection Systems

	cost	F1	Acc
ltfdf	**0.1345**	**0.74**	**0.88**
ltfidfdf	**0.1290**	**0.73**	**0.88**
ltfcnet	0.1367	0.73	0.87
tf	0.1295	0.72	0.88
ltfdfcnet	0.1515	0.71	0.86
ltfidfdfcnet	0.1350	0.71	0.87
tfidf	0.1333	0.71	0.87
tfifddf	0.1496	0.70	0.86

Modified Fractional	0.1370	0.70	0.87
tfdf	0.1540	0.69	0.86
ltfidfcnet	0.1481	0.69	0.86
Cosine Similarity	0.1505	0.69	0.86

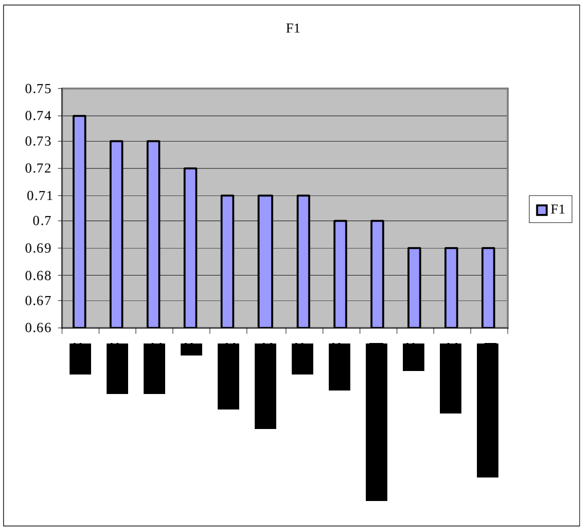

Figure 1. Performance of Link Detection Systems using F1-Measure

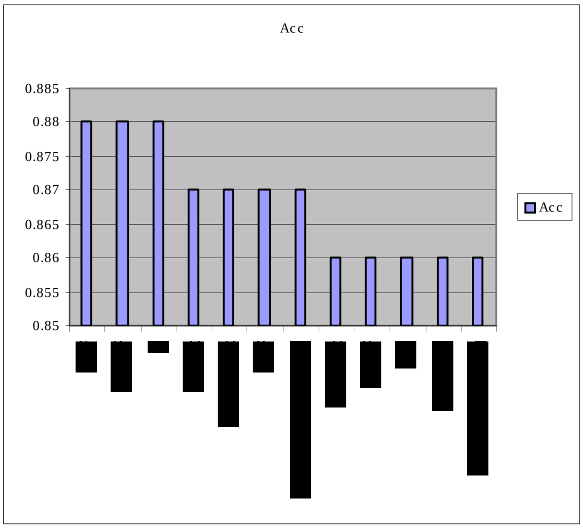

Figure 2. Performance of Link Detection Systems using Accuracy

Performance of Linear tf-df is better than all the other methods evident from table 2 and figure 1-3. Next best performance is achieved by linear tf-idf-df. Our observation is that the inclusion of idf factor reduces the false positives as well as the true positives. Figure 3 shows the comparison of true positive (tp), false positive (fp), true negative (tn) and false negative (fn) values of linear tf-df and linear tf-idf-df. It is clear that inclusion of idf factor reduces the true positive that in turn reduces the F1-measure. However it increases the true negatives (i.e. reduces the false positives) that help linear tf-idf-df to maintain the accuracy high. Low false positives of the linear-tf-idf-df helps to reduce the cost lesser than linear tf-df.

Next better F1- measure is achieved by linear tf-cnet. Figure 4 shows the comparison of true positive, false positive, true negative and false negative values of linear tf-df and linear tf-cnet. As linear tf-cnet has achieved less true positives than the linear tf-df, its F1-measure and accuracy have been reduced.

Tf as simple form achieves less true positives, but manages to achieve less false positives. Hence increases true negatives. This leads to better accuracy value. However due to low true positive values its f1- measure has reduced. Linear tf-df-cnet (ltfdfcnet) achieves more number of true positives as well as false positives. Due to its high false positive its F1- measure and accuracy are low and cost is very high.

Other models like linear tf-idf-df-cnet (ltfidfdfcnet) and tf*idf performed better than the basic cosine similarity method. tf*df, linear tf-idf-cnet (ltfidfcnet) performed on par with cosine similarity measure.

Comparing models without query expansion, model that uses modified fractional similarity performs better than cosine similarity. Results show that using modified fractional similarity reduces the false positive and improves the performance.

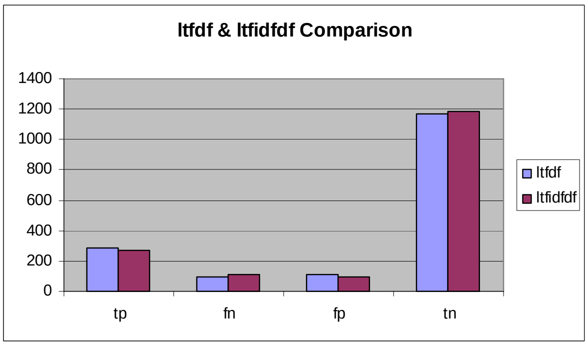

Figure 3. Comparison of Linear tf-df and Linear tf-idf-df

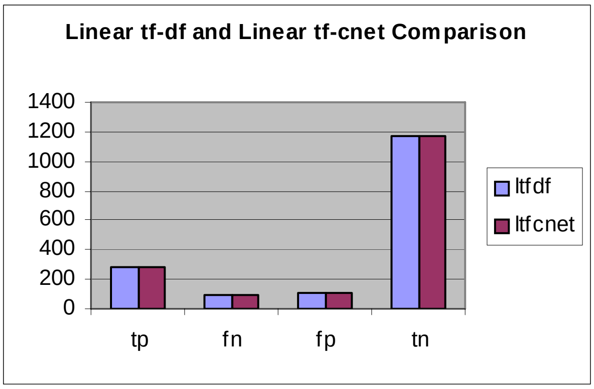

Figure 4. Comparison of Linear tf-df and Linear tf-cnet

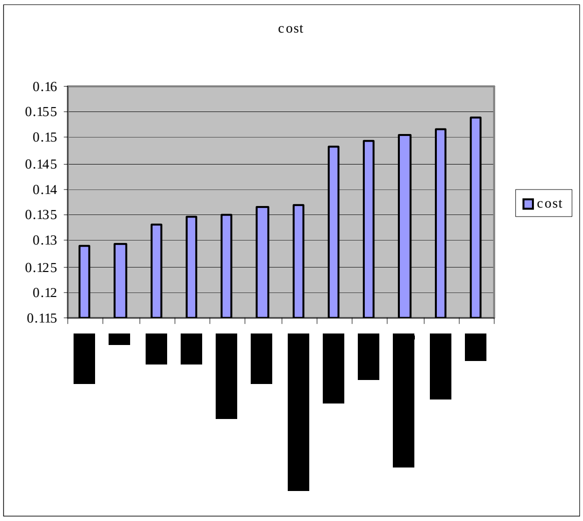

Figure 5. Performance of Link Detection Systems using cost

In link detection system low cost indicates better performance. Systems in link detection are expected to produce less false positives while maintaining high true positives. Hence cost function is designed in favor of less false positives. From the figure 5 we can observe that lowest cost is achieved by linear tf-idf-df (ltfidfdf). Other systems like tf, generative tfidf, linear tf-df (ltfdf), and linear tf-idf-df-cnet (ltfidfdfcnet) have achieved a comparable performance.

5. Conclusions and Enhancement

In this work we have proposed link detection systems with and without query expansion techniques. Among system without query expansion technique, modified fractional similarity shows better performance than the basic cosine similarity model.

Systems with query expansion technique use social cohesion as the basis for preparing the Cohesion-Model. We have taken four factors of social cohesion and constructed number of link detection system with various combinations of the basic factors sense of belonging (tf), sense of moral, goal consensus (df) and network cohesion (cnet). Most of the models constructed with various combination of social cohesion factor performed better than base cosine similarity method.

Best performance is achieved by linear combination of tf and df. It is able to produce high F1-measure and accuracy with low cost. Surprisingly linear combination of all four factors i.e. tf, idf, df and cnet didn't produce good results because of its low true positives.

In the proposed link detection system all the social cohesion parameters are weighted equally. Use of different weight to different parameter didn't show much difference in the output (not shown in experimental results).

A Link Detection system should be capable of dealing with stories in multiple languages. In this current research we have taken care of documents in English language only. It will be an interesting extension to find the performance of this model in other (regional) languages. Investigating the impact of change in language style will be another interesting study. In our future research we plan to improve the method of term weighting. Various feature selection methods have to be explored in future.

References

[1] Allan, J., "Introduction to Topic Detection and Tracking,Topic Detection and Tracking: Event-based Information Information Organization", Kluwer Academic Publishers, pp. 1-16, 2002.

[2] Allan, J. Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang., "Topic detection and tracking pilot study: Final report". In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, San Francisco, CA,. Morgan Kaufmann publishers, Inc., pp. 194-218, 1998.

[3] Topic detection and tracking (tdt) project.homepage:http://www.nist.gov/speech/tests/tdt/.

[4] Francine Chen, Ayman Farahat, Thorsten Brants, "Multiple Similarity Measures and Source-Pair Information in Story Link Detection", In proceedings of HLT-NAACL pp. 313-320, 2004.

[5] Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., and Thomas, S., "Relevance models for topic detection and tracking", In Proceedings of Human Language Technologies Conference, HLT, pp. 104-110,2002.

[6] Victor Lavrenko, "A Generative Theory of Relevance", PhD Thesis, University Of Massachusetts Amherst, September 2004.

[7] Yang, Y., Ault, T., Pierce, T., and Lattimer, C. W., "Improving text categorization methods for event tracking", In Proceedings of the 23rd Annual international ACM SIGIR Conference on Research and Development in

information Retrieval (Athens, Greece, July 24 - 28,). SIGIR '00. ACM Press, New York, NY, pp. 65-72, 2000.

[8] Ayman Farahat, Francine Chen, Thorsten Brants, “Optimizing Story Link Detection is not Equivalent toOptimizing New Event Detection”, In proceedings of ACL, pp. 232-239,2003.

[9] Ramesh Nallapati and James Allan, “Capturing Term Dependencies using a Language Model based on Sentence Trees”, CIKM’02, November 4–9, McLean, Virginia, USA. 2002.

[10] Lakshmi, K. Mukherjee, S., “An Improved Feature Selection using Maximized Signal to Noise Ratio Technique for TC”, In proceedings of Information Technology: New Generations, 2006. ITNG 2006, pp. 541 – 546, April 2006,.

[11] J. Allan, V. Lavrenko, D. Frey, and V. Khandelwal, "UMass at TDT 2000", Proceedings of the Topic Detection and Tracking Workshop, 2000.

[12] Maria Elena Figueroa,D. Lawrence Kincaid, Manju Rani,Gary Lewis, “Communication for Social Change: An Integrated Model for Measuring the Process and Its Outcomes”, The Rockefeller Foundation, New York, 2002.

[13] Hema Raghavan and James Allan, “Using soundex codes for indexing names in asr documents”, In Proceedings of the HLT NAACL Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval, 2004.