Improving Diversity of Focused Summaries Through The Negative Endorsements of Redundant Facts

Palakorn Achananuparp[1], Xiaohua Hu[1], Lifan Guo[1], Tingting He[2], Yuan An[1], and Zhoujun Li[3]
[1]*College of Information Science and Technology, Drexel University, Philadelphia PA, USA*
[2]*Department of Computer Science, Central China Normal University, Wuhan, China*
[3]*School of Computer Science and Engineering, Beihang University, Beijing, China*
pkorn@drexel.edu, thu@cis.drexel.edu, lifan.guo@ischool.drexel.edu, the@mail.ccnu.edu.cn,
yuan.an@ischool.drexel.edu, lizj@buaa.edu.cn

Abstract

We present NegativeRank, a novel graph-based sentence ranking model to improve the diversity of focused summary by performing random walks over sentence graph with negative edge weights. Unlike the typical eigenvector centrality ranking, our method models the redundancy among sentence nodes as the negative edges. The negative edges can be thought of as the propagation of disapproval votes which can be used to penalize redundant sentences. As the iterative process continues, the initial ranking score of a given node will be adjusted according to a long-term negative endorsement from other sentence nodes. The evaluation results confirm that our proposed method is very effective in improving the diversity of the focused summary, compared to several well-known text summarization methods.

1. Introduction

Query-focused summarization is a specific text summarization task which aims to create a short list of facts that responds to a particular information need. Several extractive approaches have been proposed to select the representative sentences containing vital facts which pertain to a given query. Compared to the generic summarization task, query-focused summarization is considered a more complex task due to various reasons. First, many extractive methods used in generic summarization rely on finding a set of sentences which represent the overall theme of a document collection. However, the candidate sentences of the focused summary do not necessarily reflect the major topics of the corpus. Moreover, although a size of a focused summary is somewhat arbitrarily defined, it is still relatively short compared to a typical text document. Because of the limited size, it has been argued that each representative sentence in a focused summary should consist of at least two key properties: *saliency* and *novelty* [17][18] [31]. The first property dictates that a representative sentence should contain vital facts relevant to the specific information need. Next, the novelty property is defined to ensure that those facts should be a unique with respect to one another. In other words, the whole extracted summary should be highly diverse in its factual coverage of relevant information.

In this paper, we introduce an eigenvector centrality-based ranking model called *NegativeRank* to diversify the factual coverage of focused summary. Our goal is to improve diversity of focused summary from factual coverage aspect. To achieve that, we define a task of maximizing the diversity of a set of summaries as maximizing the novelty (or minimizing the redundancy) of the individual sentence. The main contribution of the proposed method is in the application of the negative edge weights to reduce redundancy among representative sentences. Given a specific summary topic, we represent a set of relevant sentences as a set of vertices whose edge weights indicate the degree of similarity between sentences. Next, a negative sign is assigned to the edge weights in order to model the redundancy relations between sentences. Then, we re-rank a score of each sentence based on its long-term negative endorsement induced through random walks. The next contribution is in the use of sentence semantics to improve edge weighting of the sentence graphs. Specifically, we aim to address the issue of natural language variation which significantly affects the similarity judgment. In order to provide an accurate measure of inter-sentence similarity, we employ sentence semantic similarity to compute the semantic similarity of sentences based on the comparison between meaningful constituents in the sentence.

The paper is organized as follows. First, we describe the related work in section 2. Next, we explain the proposed methods in section 3. In section 4, we outline the experimental evaluation, including data sets, evaluation metrics, and procedures employed in

the study. Finally, we discuss the results and conclude the paper in section 5 and 6, respectively.

2. Related Work

In recent years, text summarization research has begun to shift its focus from generic summarization task [14][20][21] to query-focused summarization where the summary is generated from multiple sources to respond to a specific information need. In general, this task is parallel to sentence retrieval task in information retrieval. Many methods [22] [26] [28] have been proposed to generate the focused summaries where the summarization problem is formulated as sentence extraction task.

Diversity issue is one of the major concerns in both generic and focused summary. There are a growing amount of works [10][17][18][28][31] which try to integrate diversity into the sentence ranking function itself. For example, Zhu et al. propose a unified ranking algorithm called GRASSHOPPER which is based on random walks over an absorbing Markov chain. The representative sentences which have been selected into the summary become absorbing states, effectively transforming their transition probabilities to zero. The absorbing nodes will drag down the scores of the adjacent nodes as the walk gets absorbed. On the other hand, the nodes which are far away from the absorbing nodes still get visited by the random walk.

In a bigger context, the diversity issue is one of the most important topics in several research areas. Perhaps, the most well-known work is Maximal Marginal Relevance (MMR) [9] in which redundancy reduction method is first introduced to rerank the search results. Since then, it has become the most commonly used method to reduce redundancy in text summarization. Subsequent works in information retrieval research attempt to establish a theoretical framework of diversity ranking and evaluation [3][11] [30]. Our method differs from other ranking methods in the application of random walks over the negative-edge graph. Most graph-based ranking models [10][22][31] are inspired by the PageRank algorithm [8]. Therefore, they employ eigenvector centrality to measure the importance of nodes in sentence graph. Under this model, a node is considered to be important if it is linked to other important nodes. Simply, it receives a high recommendation vote from the adjacent nodes. In contrast, our strategy is to focus on reranking the initial scores by utilizing the negative edges to model the negative endorsements. Redundant nodes are those which receive a significant number of disapproval votes.

Next, the applications of negative edges in ranking model have been explored in other areas, such as trust ranking [13], social network mining [15], and complex question answering [1]. For example, de Kerchove et al. [13] propose the PageTrust algorithm as an extension to the original PageRank algorithm by including negative links as the propagation of distrust among web pages. Their method ranks the nodes using both positive and negative links. Similarly, Kunegis et al. [15] define an eigenvector ranking method called signed spectral ranking which considers both positive and negative links to model friend and foe relationships in the social network. On the other hand, our method only considers the negative links to model redundancy relationship among sentences. Finally, the semantic structure of sentence has been applied in a few text mining applications. In text categorization, Shehata et al. [25] propose conceptual term frequency as a new term weight scheme computing at sentence semantic level. Our motivation to measure sentence similarity at sentence semantic level is similar to [29]. However, their similarity formulation is relatively simple and does not consider the weighted influence of the different sentence constituents.

3. The Proposed Methods

3.1. The NegativeRank Model

We define the task of maximizing the diversity of a set of summaries as maximizing the novelty or minimizing the redundancy of the individual sentences. To that end, the proposed method focuses on two key properties of a focused summary: saliency and novelty, previously described in section 1. To incorporate these properties, we define two relations represented by two types of edges. First, the positive edges denote the relevance between sentence nodes and the summary topic. On the other hand, the negative edges represent the redundancy between sentence nodes. Intuitively, the negative-signed edge can be interpreted as a disapproval vote between nodes as opposed to a recommendation vote of the normal positive-signed edge. The absolute value of negative edge weight represents the degree of similarity between sentences. Thus, given the two components, the relevance structure indicates how salient a sentence node is with respect to a given topic while the redundancy structure indicates how redundant a given node is compared to other nodes. Then, we perform random walks over the negative-edge graph to find a long-term negative endorsement of each sentence node. In this case, the stationary distribution, derived at the end of Markov Chain, serves as a re-ranked score of each sentence.

Given a summary topic *q* and a set of relevant sentences, we first define *G*= (*V*,*E*) as an undirected graph where V is a set of vertices representing *n* sentences, E is a set of edges representing the similarity between vertices where $E \subset V \times V$. We can represent graph *G* as an *n* x *n* weighted matrix *S* where S_{ij} is a non-negative similarity score sim(*i*,*j*) of node *i* and *j*. If *i* and *j* are unrelated, then S_{ij} = 0. From *S*, we can derive an *n* x *n* normalized similarity matrix *A* such that each element A_{ij} in *A* is the normalized value of S_{ij} such that $A_{ij} = \frac{S_{ij}}{\sum_{k=1}^{n} S_{ik}}$ and all rows in *A* sum to 1.

Next, given a specific summary topic *q*, we define a vector *r* where each element r_i is a relevance score *rel*(*i*,*q*) between sentence *i* and topic *q*. Then, we transform *r* into an *n* x *n* normalized relevance matrix *B* from the outer product of an all-1 vector and r^{T} such that each element $B_{ij} = \frac{r_i}{\sum_{k=1}^{n} r_k}$ and all rows of *B* sum to 1. Then, given the probability matrix *A* and *B*, we define a transition matrix *P* as follow:

(3.1) $$P = dA + (1-d)B$$

where *d* is a damping factor with a real value of [0,1], *A* is a normalized similarity matrix, and *B* is a normalized relevance matrix. Since all rows in *P* have non-zero probabilities which add up to 1, *P* is a stochastic matrix where each element P_{ij} corresponds to the transition probability from state *i* to *j* in the Markov chain. Thus, *P* satisfies ergodicity properties and has a unique stationary distribution $\vec{\pi}P = \vec{\pi}$. Notice that equation 3.1 essentially defines random walks over regular sentence graph. However, the stationary distributions derived from the transition matrix *P* do not take into account the redundancy between sentence nodes.

Thus, to address this issue, we modify the original graph *G* such that all edge weights in *G* have a negative sign. As such, we define G^{-} = (*V*,E^{-}) as an undirected graph where *V* is a set of *n* sentence vertices, E^{-} is a set of negative edges where $E^{-} \subset V \times V$. Specifically, the negative edges in G^{-} represent the degree of redundancy between nodes. Then, we define an *n* x *n* normalized sentence redundancy matrix *M* as an all-negative matrix of the sentence similarity matrix *S* where $M_{ij} = -\frac{S_{ij}}{\sum_{j=1}^{n} S_{ij}}$ and all rows of *M* sum to -1. Next, we incorporate the redundancy relation into a new transition matrix *Q* as follow.

(3.2) $$Q = dM + (1-d)cB$$

where *M* is a normalized sentence redundancy matrix and *B* is a normalized relevance matrix. To ensure that *Q* is ergodic, we multiply matrix *B* with a scaling factor *c*. The value of *c* is determined by the conditions that all elements in *Q* should be non-negative and each *i*-th row of *Q* should add up to 1. Since all rows of *M* add up to -1 and all rows of *B* add up to 1, *c* is a function of *d* where $c = \frac{1+d}{1-d}$. Since all rows in *Q* have non-zero probabilities which add up to 1, *Q* is ergodic. Thus, it has a unique stationary distribution $\vec{\pi}Q = \vec{\pi}$. Finally, we rank each node *i* according to its stationary probability $\vec{\pi}_i$. Following the matrix notation, the simplified NegativeRank equation can be written as follow:

(3.3)

$$DR^{t}(i) = (1+d)\cdot\frac{rel(i,q)}{\sum_{i=1}^{n} rel(i,q)} - d\sum_{j\in adj(i)} \frac{sim(i,j)}{\sum_{k=1}^{n} sim(j,k)} DR^{t-1}(j)$$

where *d* is a damping factor with a real value of [0,1]. Additionally, *d* serves as a penalty factor of redundancy. *rel*(*i*,*q*) is the relevance score of sentence *i* given summary topic *q*. And sim(*j*,*i*) is a similarity score of sentence *j* and *i*.

To estimate the value of *rel*(*i*,*q*), we employ a sentence weighting function described in Allen et al. [4] as it is shown to consistently outperform other relevance models at the sentence level. It defines the relevance score of sentence *s* given query *q* as a dot product between TFISF (term frequency times inverse sentence frequency) sentence vector and TF-weighted query vector. Lastly, we compute rank convergence using Kendall tau distance to determine a stopping point of NegativeRank iteration.

Note that other saliency weighting methods can also be used to supply the alternative initial ranking distribution, e.g., LDA topic model [7], topic-sensitive eigenvector centrality [22], query-likelihood language model, etc. In addition, we can adapt NegativeRank to generic summarization by replacing the relevance function with other saliency functions, e.g. word probability [21], lead-based scoring [6], and eigenvector centrality [14][20].

3.2. Sentence Semantic Similarity

A common approach to determine edge weights between two sentence nodes is by computing cosine similarity between the vector representation of the two sentences. However, it does not consider the variability of natural language expression which is particularly crucial in the sentence similarity judgment. To cope with the issue, we employ a sentence similarity measure which is based on the comparison between semantic structures of sentences. The particular structure, known as *verb-argument structure*, describes the relationships between sentence constituents, i.e.

verb and arguments, and their *semantic roles*. Typical notions used to describe the semantic roles are as follows. First, *rel* denotes a verb or relation between two or more arguments. Arg0 denotes a prototypical agent, Arg1 denotes a prototypical patient or theme of a given verb, and ArgM denotes an adjunctive argument (e.g., ArgM-LOC specifies location-related argument). Sentences which are richer in meaning may contain one or more verb-argument structures.

Based on the aforementioned structure, we define the sentence semantic similarity measure as follows. First, each sentence can be broken down into *m* verb-argument structures. Each verb-argument structure consists of a verb *r* and an *n* number of argument components. Each argument component is composed of text segment *t*. Then, given sentence *i* and *j*, the similarity score between verb-argument structures v_i and verb-argument structure v_j is determined by two similarity components: the verb similarity $V(r_i,r_j)$ and the argument similarity $A_k(t_i,t_j)$.

(3.4)

$$S(v_i, v_j) = \alpha \cdot V(r_i, r_j) + \frac{(1-\alpha)}{n} \cdot \sum_{k=0}^{n} A_k(t_i, t_j)$$

where α is a coefficient that controls the weight between verb similarity component and argument similarity component while *n* is a total number of argument components. Following the results from [1], we set $\alpha = 0.5$ in this work. This implies that 50% of verb-argument structure similarity comes from the verb similarity component while the rest are uniformly contributed from *n* argument similarity scores.

Verb similarity. We use a modified gloss-overlap similarity measure [5] to compute the verb similarity $V(v_i,v_j)$. Essentially, two verbs are semantically similar if they share the same meaning measured by the textual overlap between their dictionary definitions (gloss). As each word (dictionary form) can carry multiple meanings (word senses), the most similar senses are used to represent their corresponding lexical similarity. The following equations describe the similarity measure:

(3.5)
$$sim_{k,l}(r_i, r_j) = \frac{|g(k_i) \cap g(l_j)|}{|g(k_i) \cup g(l_j)|}$$

(3.6)
$$V(r_i, r_j) = \max_{k,l}[sim_{k,l}(r_i, r_j)]$$

where $sim_{k,l}(r_i,r_j)$ is the gloss-overlap similarity between a word sense *k* of verb r_i and a word sense *l* of verb r_j, $g(k_i)$ is a gloss of the word sense *k* of r_i and $g(l_j)$ is a gloss of the sense *l* of r_j. Gloss is represented as a bag of words in the calculation. Then, the verb similarity $V(r_i,r_j)$ is obtained from gloss pair that gives the maximum gloss-overlap score. To obtain glosses, we search WordNet lexical taxonomy.

Intra-argument similarity. To compute the similarity of the matching argument classes, we consider argument texts as multi-word phrases and compute the similarity between text segments of the corresponding components based on their n-gram phrasal overlap score [5][23]. The formulas are defined as follows.

(3.7)
$$Overlap(t_i, t_j) = \sum_{k=1}^{n} \sum_{m} k^2$$

(3.8)
$$A_k(t_i, t_j) = tanh(\frac{overlap(t_i, t_j)}{|t_i,| + |t_j|})$$

where *m* is a number of *i*-word phrases that appear in text segments. Equation 3.8 is a normalized form of equation 3.7 via the hyperbolic tangent function to minimize the effect of the outliers [23].

Inter-argument similarity. If an adjunctive argument ArgM is presented, we simply treat all of its subclasses, e.g. ArgM-LOC, ArgM-TMP, etc., as a single class ArgM. Then, we exhaustively compute its A_k score from all possible inter-argument comparison, such as ArgM vs. Arg0, ArgM vs. Arg1, etc. The maximum A_k score is chosen as the final score for ArgM. After that, the final ArgM score is added to the inter-argument similarity scores.

Finally, the similarity of sentence *i* and *j* is derived from the verb-argument structure pair which produces the maximum $S(v_i,v_j)$ score.

3.3. The Overall Process to Extract Summaries

3.3.1. Preprocessing. Starting from the preprocessing step, we first assign a semantic role to each sentence constituent using a semantic role labeler [12]. Next, we derive a set of verb-argument structures for each sentence based on the semantic role information. Then, we extract word features from the sentence collection by tokenizing sentences into single words, removing non content-bearing words, e.g., articles, conjunctions, prepositions, etc., and stemming the tokens using Porter Stemmer.

3.3.2 Sentence Retrieval. After preprocessing step, we use vector-space model to retrieve the relevant sentences. Free-form narrative field associated with each summary topic is used as a query. The relevance score between the sentence and query is derived from a cosine similarity between conceptual term frequency (CTF) weighted vectors of a sentence and CTF-

weighted vector of a given query. In this work, we adopt Shehata et al.'s formulation of CTF [25], in which a CTF of term *i* in sentence *j* is computed as a linear combination of its normalized term frequency and normalized conceptual term frequency. We assign single-word tokens as the conceptual term features and compute CTF_i weight for each conceptual term feature *i*. After retrieving sentences, the top-500 relevant sentences is selected from the initial retrieved set.

3.3.3 Sentence Re-ranking. The next step is to re-rank the list of relevant sentences, obtained from the previous stage, using the sentence ranking model. First, we represent the list of relevant sentences as an undirected graph with negative edges. Different edge weighting schemes are considered. In a case where sentence semantic structure similarity is used, a set of verb-argument structures of sentences are utilized as an additional input. The relevance models of the retrieved sentences and a query are formulated. The relevance sub-graph is represented by the positive edges between the query node and the sentence nodes. After the ranking scores are calculated, the top-*k* representative sentences are selected as the summary. The summary length is cut off at 250 words.

We estimate the parameters of the NegativeRank model on DUC06's task 1 through 5 (10% of DUC06 tasks). The training set contains 125 documents and approximately 3,400 sentences. The optimal parameter settings for NegativeRank are $d = 0.8$ and $c = 9$. The thresholds for inter-sentence similarity score for *SS*, *TFIDF*, and *JAC* are set to 0.4, 0.2, and 0.1, respectively.

4. Experimental Evaluation

4.1. Data Sets

We conduct a query-focused summarization evaluation using the DUC 2006 (DUC06) and DUC 2007 (DUC07) data sets. These publicly-available data sets are prepared by human experts at NIST to be used in Document Understanding Conferences for evaluating document summarization systems. Each data set comprises a set of topics (50 topics for DUC06 and 45 topics for DUC07), a set of 25 relevant news articles, and a set of human-extracted summaries for each topic to be used as the reference. Each topic contains a title and a brief narrative. The main task is to generate a 250-word summary corresponding to each summary topic description.

4.2. Evaluation Metrics

Table 1. Summary of the NegativeRank variants

Abbreviation	Relevance Score	Inter-Sentence Similarity
SB+SS	SumBasic	Sentence-level structural similarity
SB+TFIDF	SumBasic	TFIDF-weighted cosine similarity
SB+JAC	SumBasic	Jaccard coefficient
REL+SS	TFISF	Sentence-level structural similarity
REL+TFIDF	TFISF	TFIDF-weighted cosine similarity
REL+JAC	TFISF	Jaccard coefficient
1+SS	Uniform distribution	Sentence-level structural similarity
1+TFIDF	Uniform distribution	TFIDF-weighted cosine similarity
1+JAC	Uniform distribution	Jaccard coefficient

We adopt three evaluation metrics normally employed in multi-document summarization evaluation. These are ROUGE-2 (R-2), ROUGE-SU4 (R-SU4), and Basic Elements (BEs). Basically, ROUGE score is computed from a lexical n-gram recall between system-extracted summaries and human-constructed reference summaries. Since ROUGE evaluates the quality of summaries solely on the surface-level overlap, we also compute Basic Elements score (BEs) [16] to compare the overlap between the minimal-length semantic units (Basic Element). ROUGE package [19] and BEwT-E package [27] are used to compute ROUGE scores and BE scores, respectively. Due to space limitation, more detail on BEwT-E scoring formula is described in [27]. Because human-constructed summaries are used as the gold standard to evaluate the quality of the extracted summaries, the upper-bound scores of each data set can be derived by computing R-2, R-SU4, and BE scores between the reference summaries.

4.3. Methods to Compare

We compare the performance of the proposed method with several well-known baseline methods, including SumBasic [21], Maximal Marginal Relevance (MMR) [9], Topic-Sensitive LexRank [22]. In addition, we also use an inverse ranking of LexRank scores (*LexRankInv*) as a direct comparison to the proposed method. In particular, we are interested to see whether using a simple backward ranking of LexRank scores will yield the same results as employing the negative endorsements in extracting a diversified summary. To extract the representative sentences for *LexRankInv*, we run the topic-sensitive LexRank algorithm to find the stationary distribution for each sentence node. However, representative sentences are

Table 2. ROUGE and BE scores of the best NegativeRank variant and the baseline methods. The upper bounds are derived from comparing benchmark summaries against other benchmark summaries.

Method	DUC06			DUC07		
	R-2	R-SU4	BE	R-2	R-SU4	BE
Human Average	*0.1125*	*0.1710*	*0.2349*	*0.1410*	*0.1916*	*0.2600*
SB	0.0659	0.1225	0.1456	0.0852	0.1389	0.1771
MMR	0.0757	0.1308	0.1444	0.0915	0.1420	0.1581
LexRank	0.0785	**0.1394**	0.1597	0.0967	0.1528	0.1779
LexRankInv	0.0555	0.1126	0.1211	0.0699	0.1260	0.1423
NegativeRank	**0.0789**	0.1341	**0.1609**	**0.1017**	**0.1535**	**0.1781**

Table 3. ROUGE and BE scores of different NegativeRank variants. The best results are in bold.

Variant	DUC06			DUC07		
	R-2	R-SU4	BE	R-2	R-SU4	BE
SB+SS	0.0729	0.1302	0.1496	0.0904	0.1441	0.1719
SB+TFIDF	0.0624	0.1198	0.1381	0.0825	0.1368	0.1614
SB+JAC	0.0606	0.1187	0.1315	0.0775	0.1307	0.1589
REL+SS	**0.0789**	**0.1341**	**0.1609**	**0.1017**	**0.1535**	**0.1781**
REL+TFIDF	0.0781	0.1336	0.1541	0.0973	0.1533	0.1758
REL+JAC	0.0762	0.1315	0.1268	0.0962	0.1496	0.1742
1+SS	0.0728	0.1298	0.1517	0.0950	0.1500	0.1765
1+TFIDF	0.0677	0.1240	0.1400	0.0883	0.1413	0.1643
1+JAC	0.0667	0.1231	0.1390	0.0901	0.1444	0.1674

ranked in ascending order according to its stationary distribution instead of descending order.

Next, several NegativeRank variants are defined based on the combinations of the initial ranking distribution: SumBasic (*SB*), TFISF-based relevance function (*REL*), and a uniform distribution 1/*n*, and inter-sentence similarity measure: sentence semantic similarity(*SS*) described in section 3.2, TFIDF-weighted cosine similarity (*TFIDF*), and Jaccard coefficient (*JAC*). The summary of NegativeRank variants is shown in table 1. For each variant, we use the following notion: *relevance model + sentence similarity model*, to describe the underlying methods used in the variant model in an abbreviated form.

5. Results and Discussion

Table 2 and 3 display the average R-2, R-SU4, and BE scores evaluated of the baselines and NegativeRank variants, respectively. Overall, the best NegativeRank variant consistently outperforms other baselines across all three evaluation metrics. Considering all individual variants, *REL+SS* is the best performer on both data sets. The best average R-2, R-4, and BE scores for DUC06 tasks are 0.0789, 0.1394, and 0.1609, respectively, while the best average R-2, R-4, and BE scores DUC07 tasks are 0.1017, 0.1535, and 0.1781, respectively. The regular eigenvector centrality method such as *LexRank* also produces the highly competitive results, compared to *NegativeRank*'s. There is one instance where *LexRank* slightly outperforms *NegativeRank* on R-SU4 metric, but the difference is not statistically significant. Moreover, *LexRankInv* produces significantly inferior scores than most *NegativeRank* variants. This suggests that the negative endorsement is not merely a backward ranking of the regular eigenvector centrality.

The results confirm our expectation that the methods which consider both the relevance and the novelty should produce a better focused summary than the novelty-centric methods. By supplying the relevance scores as the initial ranking probabilities, the sentence ranking model produces the best results. For example, the performance scores of 1+*SS*, 1+*TFIDF*, and 1+*JAC* are significantly lower, $p<0.05$, than their corresponding counterparts, e.g. *REL+SS*, REL+TFIDF, and *REL+JAC*. In addition, the summaries obtained from NegativeRank, LexRank, and inverse LexRank suggest the effectiveness of our method. For instance, task# D0706 requires the summary to focus on the main events and important personalities in Myanmar surrounding the government changed in 1988. The reference summary created by

the human expert contains six distinct facts. In this instance, the focused summary obtained from NegativeRank only misses one fact while the summary generated by LexRank misses two facts. Moreover, the first two sentences in LexRank's summary are redundant while the summary obtained from inverse LexRank does not contain any relevant facts (see appendix).

Next, methods which employ sentence semantic similarity measure consistently outperform other variants across all evaluation metrics. For example, the scores of *SB+SS* are significantly higher than those of *SB+TFIDF* and *SB+JAC*, $p<0.05$. The similar results can be seen in the cases of *REL+SS* and *1+SS*. This suggests that the application of sentence semantic structure in edge weighting provides a significant contribution to redundancy reduction among nodes in the sentence graphs.

6. Conclusions and Future Work

We propose a graph-based sentence ranking model to extract the representative sentences for query-focused summary. The major contributions of our work are as follows. First, our model extract the novel sentences through random walks over a negative-edge graph. Second, to overcome a shortcoming of a cosine similarity based similarity measure, we utilize the sentence semantic structure to deal with the issue of natural language variation when comparing the similarity between sentences. The experimental results show that the proposed method outperforms many existing ranking models. Several directions for the future work are considered. First, we plan to extend the evaluation of NegativeRank to other related applications. For example, NegativeRank can be used explore the diversity ranking in social network mining. Furthermore, we plan conduct a more comprehensive evaluation of NegativeRank with respect to other state-of-the-art ranking models [31][10].

Acknowledgement

This research work is supported in part from the NSF Career grant IIS 0448023, NSF CCF 0905291, NSF IIP 0934197, NSFC 09020005 "Chinese Language Semantic Knowledge Acquisition and Semantic Computational Model Study," and the Program of Introducing Talents of Discipline to Universities B07042 (China).

7. References

1. Achananuparp, P., Yang, C.C., and Chen, X. (2009) Using Negative Voting to Diversify Answers in Non-Factoid Question Answering. In Proc. of CIKM 2009, Hong Kong.
2. Achananuparp, P., Hu, X., and Yang, C.C. (2009) Addressing the Variability of Natural Language Expression in Sentence Similarity with Semantic Structure of the Sentences. In Proc. of PAKDD 2009, Bangkok, 548-555.
3. Agrawal, R., Gollapaudi, S., Halverson, A., and Ieong, S. (2009) Diversifying Search Results. In Proc. of WSDM'09, 5-14.
4. Allan, J., Wade, C., and Bolivar, A. (2003) Retrieval and novelty detection at the sentence level. In Proc. of SIGIR '03, ACM, New York, NY, 314-321.
5. Banerjee, S., and Pedersen, T. (2003). Extended gloss overlap as a measure of semantic relatedness. In Proc. of IJCAI'03, Acapulco, 805-810.
6. Brandow, R., Mitze, K., and Rau, L.F. (1995) Automatic condensation of electronic publications by sentence selection. Inf. Process. Manage., 31(5), 675-685.
7. Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003) Latent dirichlet allocation. Journal of Machine Learning Research, 993-1022.
8. Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7).
9. Carbonell, J. and Goldstein, J. (1998) The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In Proc. of SIGIR'98, 335–336.
10. Chen, S.Y., Huang, M.L., and Lu, Z.Y. (2009) Summarizing Documents by Measuring the Importance of a Subset of Vertices within a Graph. In Proc. of WI 2009.
11. Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008) Novelty and diversity in information retrieval evaluation. In Proc. of SIGIR'08, 659-666.
12. Collobert, R. and Weston, J. (2007) Fast Semantic Extraction Using a Novel Neural Network Architecture. In Proceedings of ACL 2007, Prague, Czech Republic, June 23–30.
13. de Kerchove, C., and Dooren, P.V. (2008) The PageTrust algorithm: how to rank web pages when negative links are allowed? In Proc. SDM 2008 2008, 346-352.
14. Erkan, G. and Radev, D. (2004) LexPageRank: Prestige in multi-document text summarization. In Proc. of EMNLP 2004.
15. Kunegis, J., Lommatzsch, A., and Bauckhage, C. (2009) The Slashdot zoo: Mining a social network with negative edges. In Proc. of WWW 2009, 741-750.
16. Hovy, E.H., Lin, C.Y., Zhou, L., and Fukumoto, J. (2006) Automated Summarization Evaluation with Basic Elements. In Proceedings of LREC. Genoa, Italy.

17. Huang, M.L. and Chen, S.Y. (2009) Finding representative and diverse vertices within graphs. A technical report, Tsinghua University.
18. Li, L., Xue, G.R., Zha, H., and Yu, Y. (2009) Enhancing Diversity, Coverage and Balance for Summarization through Structure Learning. In Proc. of WWW 2009, 71-80.
19. Lin, C.Y. and Hovy, E.H. (2003) Automatic Evaluation of Summaries using n-Gram Co-occurrence Statistics. In Proceedings of the HLT2003 conference.
20. Mihalcea, R. and Tarau, P. (2004). TextRank: bringing order into texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain.
21. Nenkova, A. and Vanderwende, L. (2005) The impact of frequency on summarization. MSR-TR-2005-101.
22. Otterbacher, Erkan, G., and Radev, D.R. (2005) Using Random Walks for Question-focused Sentence Retrieval. In Proc. of the HLT/EMNLP 2006, Vancouver, 915-922.
23. Ponzetto, S. P. and Strube, M. (2007) Knowledge Derived From Wikipedia for Computing Semantic Relatedness, Journal of Artificial Intelligence Research, 30, 181-212.
24. Pradhan, S., Ward, W., Hacioglu, K., Martin, J.H., and Jurafsky, D. (2004) Shallow Semantic Parsing using Support Vector Machines, in Proceedings of HLT/NAACL-2004, Boston, MA, May 2-7.
25. Shehata, S., Karray, F., and Kamel, M. (2007) A concept-based model for enhancing text categorization. In Proceedings of KDD '07. ACM, New York, NY, 629-637.
26. Tang, J., Yao, L., and Chen, D. (2009) Multi-topic based query-oriented summarization. In Proc. of The SIAM Internation Conference on Data Mining (SDM 2009).
27. Tratz, S. and Hovy, E.H. (2008) Summarization Evaluation Using Transformed Basic Elements. In Proc. of TAC-08. NIST, Gaithersburg, MD.
28. Wan, X., Yang, J., and Xiao, J. (2006) Using Cross-Document Random Walks for Topic-Focused Multi-Document. In Proc. of 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), 1012-1018.
29. Wang, D., Li, T., Zhu, S., and Ding, C. (2008) Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization. In Proc. of SIGIR'08, July 20-24, Singapore, 307-314.
30. Zhai, C., Cohen, W.W., and Lafferty, J. (2003) Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In Proc. of SIGIR'03, 10-17.
31. Zhu, X., Goldberg, A., Van Gael, J., and Andrzejewski, D. (2007) Improving Diversity in Ranking using Absorbing Random Walks. In Proc. of NAACL-HLT 2007.

Appendix: The Examples of Summaries Extracted for DUC07 Task#06

NeagtiveRank's Summary:

- He said there are 24 refugee camps along the Myanmar-Thai border where members and their families of different anti-Myanmar government armed groups such as the All Burma Students' Democratic Front (ABSDF), Kayin National Union (KNU) and Democratic Alliance of Burma (DAB) are living and conducting military and "terrorist" training there involving foreigners.
- Suu Kyi won the Nobel Peace Prize in 1991 for her peaceful struggle for democracy against the military regime in Myanmar, also known as Burma.
- There are 42 NLD members of parliament in Myanmar's prisons, according to the All Burma Students Democratic Front, an exile group.
- The vice chairman of Myanmar opposition leader Aung San Suu Kyi's political party was threatened with arrest in a commentary in a government-run newspaper Sunday.

LexRank's Summary:

- The military has ruled Myanmar, also known as Burma, since 1962.
- Myanmar, also known as Burma, has been ruled by the military since 1962.
- The current military government came to power on Sept. 18, 1988 after brutally crushing a nationwide democracy movement.
- Suu Kyi won the Nobel Peace Prize in 1991 for her peaceful struggle for democracy against the military regime in Myanmar, also known as Burma.

Inverse LexRank's Summary:

- A high-ranking Myanmar military official said Sunday that the authorities made timely arrest of 40 persons in January, who allegedly attempted to commit terrorist acts in the country.
- Citing her personal physicians, who have visited her twice in her van outside Yangon, her eyes are turning yellow and she has low blood pressure, the party statement said.
- On Thursday, the Burma Lawyers Council, composed of exiles, called on the country's lawyers to endorse the convening of parliament.
- In the spirit of this philosophy, I present today in my capacity as chairman of the billion-dollar multinational Make a Buck at Any Cost Corp. my special report on American Business Sentiment toward Burma.