Using Negative Voting to Diversify Answers in Non-Factoid Question Answering

Palakorn Achananuparp
College of Information
Science and Technology,
Drexel University
Philadelphia PA 19104
pkorn@drexel.edu

Christopher C. Yang
College of Information
Science and Technology,
Drexel University
Philadelphia PA 19104
chris.yang@ischool.drexel.edu

Xin Chen
College of Information
Science and Technology,
Drexel University
Philadelphia PA 19104
bruce.chen@drexel.edu

Abstract

We propose a ranking model to diversify answers of non-factoid questions based on an inverse notion of graph connectivity. By representing a collection of candidate answers as a graph, we posit that novelty, a measure of diversity, is inversely proportional to answer vertices' connectivity. Hence, unlike the typical graph ranking models, which score vertices based on the degree of connectedness, our method assigns a penalty score for a candidate answer if it is strongly connected to other answers. That is, any redundant answers, indicated by a higher inter-sentence similarity, will be ranked lower than those with lower inter-sentence similarity. At the end of the ranking iterations, many redundant answers will be moved toward the bottom on the ranked list. The experimental results show that our method helps diversify answer coverage of non-factoid questions according to F-scores from nugget pyramid evaluation.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – abstracting methods

General Terms

Algorithm, Experimentation, Performance

Keywords

Answer ranking, negative voting, non-factoid question answering

1. Introduction

Unlike factoid questions, whose answer comprises a short text segment of 50 bytes or less, non-factoid questions are inherently more complex and require a paragraph-length answer. Generating answers for complex, non-factoid questions from a large sentence collection is a very challenging task. Since some answers contain more vital information to the questions while some answers are more trivial, it is crucial for question answering systems to present as many informative answers in the system response as possible. However, past research proposes that informativeness is not a sole criterion for selecting answers. For instance, [7] suggest that a good question answering system should provide a mixture of both informative and interesting answers in the system response. Moreover, another important criterion is the novelty of answers. In the process of selecting answers, some candidate answers may contain redundant information compared to others. For instance, the following statements "*George Edward Foreman is the oldest boxer to win a major heavyweight title*" and "*Foreman won a major boxing championship at age 45*" contain mostly the same information as "*George Foreman becomes the oldest world champion in boxing history.*" Thus, we approach the problem of generating answers for non-factoid questions from a diversification perspective. The focus of our work is on maximizing the informativeness and diversity of responses in a fixed-length extracted answer list. We propose a novel sentence-level ranking model called *DiverseRank* for re-ranking candidate answers by their informativeness and novelty.

2. Related Work

Extracting a ranked list of informative answers for complex non-factoid questions have been a major research topic for quite some time. Many researchers have proposed methods to find informative answers. The earliest approach, which starts from definition question answering research, is based on handcrafted lexico-syntactic patterns [4]. Apart from informativeness, other aspects of answers have been explored. For instance, Kor and Chua [7] propose a unigram language model constructed from various web snippets to find interesting answers. Despite improvements in system performance, generating a list of answers for complex questions remains a challenging task due to the fact that the answers do not easily fall into predictable semantic classes. Several related works have been done in community/collaborative question answering (CQA) domain. For example, many graph-based answer ranking models have been proposed [15][5]; many of which are inspired by HITS algorithm [6]. Nevertheless, many studies in CQA domain have focused on expert finding problem in community QA sites while answer diversification problem has not been addressed. Lastly, the main difference between our work and other graph-based sentence ranking models [13][16] is in voting strategy. Most graph-based ranking methods, e.g. LexRank [13], are based on PageRank algorithm [1], which relies on positive voting or recommendation between vertices. However, we hypothesize that novelty is inversely proportional to connectivity. Thus, our proposed method negatively votes down the score of any high-degree vertices.

CIKM'09, November 2–6, 2009, Hong Kong, China.

3. The Proposed Method

We propose the sentence ranking method called DiverseRank, which aims to address the aforementioned problems. The proposed method is motivated by the intuition that a good candidate answer should be *highly informative* and *novel* with respect to other answer sentences. Conceptually, informativeness of a sentence is proportional to a number of important words -- the more important words the sentence has, the higher its informativeness. Next, we postulate that novelty, a unit representing diversity, is inversely proportional to connectedness. Thus, the most novel sentence in the answer list should have the minimum similarity with respect to the rest of the answers. Based on this assumption, we compute DiverseRank score for an answer sentence according to its informative score penalized by the redundancy score from the neighboring sentences.

Given that $G = (V, E)$ is an undirected graph where V is a set of vertices representing each answer, E is a set of edges representing the similarity between vertices, and $E \subset V \times V$, DiverseRank score of V_i is defined as:

$$DR_k(V_i) = (1-d)DR_{k-1}(V_i) - d \sum_{V_j \in E(V_i)} \frac{1}{|E(V_j)|} DR_{k-1}(V_j) \quad (1)$$

Where d is a damping factor with a real-number value in [0,1] range. We use a standard value of $d = 0.85$ in this work. $E(V_i)$ is a set of edges that connect to V_i for a given similarity threshold. Binary discretization is performed for a given similarity threshold to determine the value of $E(V_i)$. From the above equation, DiverseRank can be conceptually viewed as a form of "inverse PageRank" where each neighboring vertex casts a negative vote for the other vertex. The greater the degree a vertex has, the lower the score it has in the end. Moreover, we postulate that the initial scores (DR_0) play a significant role in determining the final DiverseRank scores. Thus, we assign informative score as a starting value for each sentence vertex. Then, an iterative computation continues until convergence where there are no changes in the overall graph ranking. According to the evaluation on the test data sets, the average number of iterations to reach convergence for 100 vertices is 60 given d=0.85.

Equation 2 describes DiverseRank computation at the initialized stage (DR_0). As it can be seen, we replace the first component in equation 1 with the informative score measure denoted $Info(V_i)$. In this work, we propose two measures to estimate sentence informativeness; they are $Info_{IDF}$ and $Info_{REL}$.

$$DR_0(V_i) = (1-d)Info(V_i) - d \sum_{V_j \in E(V_i)} \frac{1}{|E(V_j)|} DR_{k-1}(V_j) \quad (2)$$

First, we model informativeness as a function of *term rarity*. That is, $Info_{IDF}$ computes the informative score of a sentence as a normalized sum of inverse document frequency (*idf*) of matching terms between question q and sentence s.

$$Info_{IDF}(s_i) = \frac{\sum_{w_i \in s_i} idf_{w_i}}{\sum_{w_j \in C_S} idf_{w_j}} \quad (3)$$

where $idf_{w_i}$is an inverse document frequency of term w_i while C_S is the entire sentence collection. Evidently, this measure scores answer sentences that contain greater number of rare words higher than the ones with a fewer number of rare words. Inherently, rarity is also a direct indication of diversity.

Alternatively, the informative score of a sentence can be derived from its *conceptual relevance* to a given question. We define conceptual relevance as a function of conceptual term frequency (*ctf*) of words in the sentence and words in the question; where, *ctf* is compute from a number of occurrences of a conceptual term (i.e. either single words or multi-word phrases) in verb-argument structures of a sentence [14]. It is based on the assumption that concepts that appear in a greater number of verb-argument structures contribute more to sentence meaning than those that appear in a fewer number of verb argument structures.

$$Info_{REL} = \frac{\sum_{w_i \in q,s} CTF_{w_i,q} CTF_{w_i,s}}{\sqrt{\sum_{w_i \in q} CTF_i^2}\sqrt{\sum_{w_j \in s} CTF_j^2}} \quad (4)$$

Where $CTF_{w_i,q}$ is a concept-based weight of w_i in question q while $CTF_{w_i,s}$ is a concept-based weight of w_i in question s. To compute *CTF*, we adopt Shehata et al.'s formulation [14] which defines a concept-based weight (henceforth CTF) of term i in sentence j as a linear combination of its normalized term frequency and normalized conceptual term frequency.

$$CTF_i = \frac{tf_i}{||s_j||} + \frac{ctf_i}{||t_j||} \quad (5)$$

where tf_i is a frequency of term i, ctf_i is a conceptual term frequency of term i which is determined by summing the occurrences of i in the verb-argument structures of sentence j, s_j is a term-frequency weighted vector of sentence j, and t_j is a conceptual term-frequency weighted vector of sentence j.

In a case where sentence informativeness is ignored, we simply initialize all sentence scores with a constant value 1. Thus, we define a purely diversity-based DiverseRank as follows:

$$DR_k(V_i) = (1-d) - d \sum_{V_j \in E(V_i)} \frac{1}{|E(V_j)|} DR_{k-1}(V_j) \quad (6)$$

To determine the degree of sentence vertex , we perform binary discretization for a given similarity threshold θ. Specifically, any vertex with edge weight below θ will have its edge removed. In order to handle variability of answer expression, we measure inter-sentence similarity at sentence semantics level. We fully describe our sentence-level structural similarity in section 3.2. In this work, we use $\theta = 0.4$ as it's empirically proved to be the optimal value in our previous work [1].

The overall process to generate a diversified answer passage is as follows. First, a semantic role labeler is employed to extract verb-argument structures for each sentence. Then, we use a vector-space model to retrieve a list of top-500 relevant sentences. For this, we combine text segments in both question topic and free-form narration fields into a single query. Relevance score between sentences and query is derived from a cosine similarity between CTF-weighted sentence vector and CTF-weighted query vector. Next, we construct conceptual term-document matrix where conceptual term features are taken directly from single word tokens. To extract the single-word tokens, we remove functional or non content-bearing words (articles, conjunctions, prepositions, etc.) from sentences but keep the cardinal numbers, and apply Porter Stemmer. Next, CTF weight is computed for each conceptual term feature. Finally, the relevance score of a sentence is calculated from a cosine similarity between CTF-weighted sentence vector and CTF-weighted query vector. A list of top-500 answer sentences is selected from the retrieved set.

After a list of top-500 relevant sentences is retrieved, we partition sentences into k subtopics using k-means clustering. In this case, the value of k is simply determined by dividing a fixed answer length by an average length of sentence in the corpus [10]. The cluster centroid is computed from cosine similarity of the corresponding conceptual term vectors. The best result from 10 runs is chosen. Then, we create a list of 100 candidate answers by selecting the top sentences from each cluster in a round-robin manner. That is, given $c_i \in C_k$ where C_k is a set of sentence clusters, candidate sentence i is selected from the top-scoring sentence from cluster c_i while candidate sentence i+1 is selected from cluster c_{i+1}, and so on until the top sentence in the last cluster c_k is selected. If the total number of selected sentences is still less than 100 after the k^{th} round, then we restart the process by selecting the next-top sentence from the first cluster and so on. The selection step is terminated after the candidate answer list contains 100 sentences.

Lastly, we rerank the answers obtained from the previous step using our DiverseRank method. First, we compute the informative score for each sentence in the list. Then, we construct adjacency matrix where inter-sentence similarity value is derived from the sentence-level structural similarity measure. Then, we build the sentence graph by discretizing the edges at similarity threshold $\theta = 0.4$. The reranking continues until the overall ranking reaches convergence. Finally, we terminate the process and truncate the answer list to 7,000 characters (following the standard procedure in related TREC evaluation).

4. Experimental Evaluation

4.1 Evaluation Metrics

We adopt a standard protocol used in the automatic evaluation of system-generated answers to complex questions called the "nugget pyramid" [9] to assess the performance of our proposed methods. In essence, the method calculates the system scores according to a weighted harmonic mean (F_1 measure) between nugget recall (NR) and nugget precision (NP). NR and NP are derived from summing the unigram co-occurrences between terms in each "information nugget" and terms from each system-extracted answer. Pourpre scoring script 1.1c [8] is used to compute F-scores. Equation 7-10 describes the formulas to compute NR, NP, and F score.

$$NR = \frac{r}{R} \tag{7}$$

$$\alpha = 100 \times (r + a) \tag{8}$$

$$NP = \begin{cases} 1 & \text{if } l < \alpha \\ 1 - \frac{l - \alpha}{l} & \text{otherwise} \end{cases} \tag{9}$$

$$F_\beta = \frac{(\beta^2 + 1) \times NR \times NP}{\beta^2 \times NP + NR} \tag{10}$$

4.2 Data Sets

We conduct the experiment on complex interactive question answering (ciQA) test set used in TREC 2006 ciQA task. Unlike traditional TREC question answering data, ciQA data focus exclusively on complex relationship questions. A relationship is defined as the ability of one entity to influence another, including both means to influence and the motivation for doing so [5]. This type of questions reflects the information needs generally faced by intelligence analysts, e.g. financial, movement of goods, family ties, communication pathways, organizational ties, co-location, common interests, and temporal. A total of thirty question topics and a detail description of their information need are prepared by human judges at NIST (National Institute of Standards and Technology). Overall, there are 2,320.87 answer sentences per test question on average. NIST assessors also create the answer key comprising a list of vital/okay information nuggets for each test question. On average, each question contains sixteen answer nuggets.

4.3 Methods Compared

In this work, we compare the performance of four baselines and eight specific DiverseRank variants. Table 1A and 1B summarizes all methods compared in the evaluation.

Table 1A. Summary of the baselines used in the evaluation

Abbreviation	Description
SB	SumBasic [12]
MMR	Maximal Marginal Relevance [3]
LexRank	Topic-Sensitive LexRank [13]
LexRank'	Inverted ranking of LexRank scores

Table 1B. Summary of the DiverseRank variants

Abbreviation	Initial Score	Inter-Sentence Similarity
SB+NG	SumBasic	N-gram overlap variant [1]
SB+CS	SumBasic	CTF-weighted cosine similarity variant
IDF+NG	IDF	N-gram overlap variant
IDF+CS	IDF	CTF-weighted cosine similarity variant
REL+NG	CTF-weighted cosine similarity	N-gram overlap variant
REL+CS	CTF-weighted cosine similarity	CTF-weighted cosine similarity variant
1+NG	Constant score of 1	N-gram overlap variant
1+CS	Constant score of 1	CTF-weighted cosine similarity variant

5. Results and Discussion

The pyramid F-scores of the twelve methods are shown in table 2. Overall, the best DiverseRank method, REL+CS, significantly outperform all four baselines (p-value<0.05). This suggests that our proposed method are effective in generating a diverse list of answers. For example, it outperforms SB and MMR by 24.71% and 48.30%, respectively. In addition, It also significantly outperforms LexRank at p<0.05 although the improvement is relatively smaller. Despite the fact that the notion of graph connectivity in DiverseRank is opposite to that in LexRank, it performs very competitively, or even superior to LexRank under certain conditions. Note that, despite the similar negative ranking mechanism, inverted LexRank (LexRank') produces inferior F-score to DiverseRank. This suggests that DiverseRank is not merely a backward ranking of LexRank. Within DiverseRank variants, the best methods also significantly outperform (p-value < 0.05) other DiverseRank methods.

When both informativeness and novelty are considered in DiverseRank model, it produces the best result, compared to the purely salient methods, e.g. SB, MMR, and LexRank. Moreover, evidence from the evaluation also suggests that the purely diversity-centric methods, e.g. 1+NG and 1+CS, perform poorer than the more inclusive methods. We believe that the best DiverseRank variants work well since they rank answers in a more balance manner.

Table 2. F-Scores of baselines and DiverseRank variants averaging across all questions. The best methods are in bold.

Method	Pyramid F-Score	% improvement of the best variant (REL+CS)
SB	0.2956	+24.71%
MMR	0.2486	+48.30%
Lex	0.3590	+2.69%
Lex'	0.3516	+4.82%
SB+NG	0.3433	+7.38%
SB+CS	0.3454	+6.70%
IDF+NG	0.3442	+7.00%
IDF+CS	0.3445	+7.10%
REL+NG	0.3400	+8.40%
REL+CS	**0.3686**	-
1+NG	0.3456	+6.67%
1+CS	0.3439	+7.18%

6. Conclusions

In this paper, we propose a novel ranking model to re-rank candidate answers of non-factoid questions using graph connectivity. Unlike the tradition graph ranking models, our proposed method employs a negative voting strategy to iteratively adjust answer score by its degree of redundancy with other candidate answers. As a consequence, answers which are highly informative but too similar to others will be iteratively voted down. The final outcome is an answer passage which has a good balance of informativeness and novelty. We conduct an empirical evaluation of our proposed method on complex question answering (ciQA) 2006 data. The results show that our DiverseRank method outperforms most baseline methods. Moreover, it performs very competitive against an effective graph ranking model, such as topic-sensitive LexRank.

Acknowledgement

This work is supported in part by NSF Career grant (NSF IIS 0448023), NSF CCF 0514679, PA Dept of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196) and PA Dept of Health Grant (No. 239667).

7. References

[1] 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, Thailand, 548-555

[2] Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7).

[3] Carbonell, J. and Goldstein, J. (1998) The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In Proceedings of SIGIR'98, 335–336.

[4] Harabagiu, S., Moldovan, D., Clark, C., Bowden, M., Hickl, A., and Wang, P. (2005) Employing two question answering systems at trec 2005. In Proc. of TREC '05, Gaithersburg, Maryland, 2005.

[5] Jurczyk, P. and Agichtein, E. (2007) Discovering authorities in question answer communities by using link analysis, In Proc. of CIKM 2007, November 06-10, 2007, Lisbon, Portugal.

[6] Kleinberg, J.M. (1999) Authoritative sources in a hyperlinked environment, Journal of the ACM (JACM), 46(5), 604-632.

[7] Kor, K. and Chua, T. (2007) Interesting nuggets and their impact on definitional question answering. In Proc. of SIGIR '07, Amsterdam, the Netherlands, 335-342.

[8] Lin, J., and D., Demner-Fushman (2005) Automatically Evaluating Answers to Definition Questions. In Proc. of HLT/EMNLP, Vancouver, 931-938

[9] Lin, J., and D., Demner-Fushman (2006) Will Pyramids Built of Nuggets Topple Over? In Proc. of the HLT/NAACL 2006, 383-390.

[10] Liu, D., He, Y., Ji, D., and Yang, H. (2006) Multi-Document Summarization Based on BE-Vector Clustering. In Proceedings of CICLing 2006, 470-479.

[11] Liu, Y., Bian, J., and Agichtein, E. (2008) Predicting Information Seeker Satisfaction in Community Question Answering. In Proc. of SIGIR'08, Singapore, July 20-24.

[12] Nenkova, A. and Vanderwende, L. (2005) The impact of frequency on summarization. MSR-TR-2005-101.

[13] 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.

[14] 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.

[15] Suryanto, M. A., Lim, E. P., Sun, A., and Chiang, R. H. (2009) Quality-aware collaborative question answering: methods and evaluation. In Proc. of WSDM '09. Barcelona, Spain, 142-151.

[16] Wan, X. and Yang, J. (2007) Towards a Unified Approach Based on Affinity Graph to Various Multi-document Summarizations. In Proceedings of ECDL 2007, 297-308