Answer Diversification for Complex Question Answering on the Web

Palakorn Achananuparp[1], Xiaohua Hu[1], Tingting He[2],
Christopher C. Yang[1], and Yuan An[1]

[1]College of Information Science and Technology, Drexel University, Philadelphia PA
[2]Department of Computer Science, Central China Normal University, Wuhan, China
pkorn@drexel.edu, thu@cis.drexel.edu, the@mail.ccnu.edu.cn,
chris.yang@ischool.drexel.edu, yuan.an@ischool.drexel.edu

Abstract. We present a novel graph ranking model to extract a diverse set of answers for complex questions via random walks over a negative-edge graph. We assign a negative sign to edge weights in an answer graph to model the redundancy relation among the answer nodes. Negative edges can be thought of as the propagation of negative endorsements or disapprovals which is used to penalize factual redundancy. As the ranking proceeds, the initial score of the answer node, given by its relevancy to the specific question, will be adjusted according to a long-term negative endorsement from other answer nodes. We empirically evaluate the effectiveness of our method by conducting a comprehensive experiment on two distinct complex question answering data sets.

Keywords: Answer diversification, answer reranking, random walk, negative-edge graph, complex question answering

1. Introduction

Automatically generating a set of answers for *complex questions*, e.g. definition, opinion, and online community questions, remains a challenging task for several reasons. First, the information needs underlying this type of questions are often subjective and ill-defined [15]. Hence, it requires one or more answer passages to generate a complete response to complex questions. Furthermore, the answers themselves do not easily fall into predictable semantic classes [10]. So, name-entity style answer extraction techniques are not likely to be effective. Moreover, it is more desirable for the automatic response to return a set of factually diverse answers.

Suppose that there are two sets of answers, *A* and *B*, generated by two different systems that response to a given question "*what effect does steroid use have on athletes' performance*?" Set *A* consists of two answers {*steroid helps boost athletic performance by improving muscle mass*, *steroids can cause many harmful effects*} while set *B* contains {*steroids enhance athletic performance*, *athletes use steroids to improve their performance*}. An information seeker would find answers in set *A* to be more useful than those in set *B*. As illustrated, the two facts in the first set are

relatively more novel than those in the second set. In other words, the factual coverage of the second set is less diverse than the first one's.

1.1 Contributions

In this research, we utilize a graph topology to rerank the answers according to their relevance and novelty. The summary of our contributions is as follows:

1. We propose a graph ranking model called *DiverseRank* to extract a diverse set of answers for complex questions. Our method is motivated by the ideas that a good answer set should contain facts which are highly relevant as well as novel. The main contribution of our work is in the use of a graph topology to reduce redundancy among answers. We represent a set of answers as a set of vertices whose edges correspond to the similarity between answer nodes. A negative sign is assigned to the edge weights to model the redundancy relationship between nodes. Then, the final ranking score for each answer node is derived from its long-term negative endorsement.

2. We conduct a comprehensive experiment on two distinct question answering data sets: a subset of Yahoo! Answers data and TREC 2006's complex questions data, to evaluate the performance of the proposed method. To measure the quality of the extracted answers, we use a *nugget pyramid* [14] evaluation and a recall-by-length curve as the performance metrics. Specifically, we measure diversity in terms the amount of common *information nuggets*, a small text fragment that describes a certain fact about a given question, between the extracted sets and the gold standard set.

1.2 Paper Organization

The rest of the paper is organized as follows. First, we review related work in section 2. Next, we describe the proposed method in section 3. In section 4, we present the experimental evaluation, including data sets, evaluation metrics, and procedures. Finally, we discuss about the results and conclude the paper in section 5 and 6, respectively.

2. Related Work

Several issues in web community question answering have been investigated, e.g. finding high-quality answers [9][19], maximizing the facet coverage [16], etc. However, answer diversity issue has not been explored as much. Diversity in ranking has long been one of the major issues in many research areas, such as text summarization and information retrieval. Many information retrieval researchers have attempted to establish several theoretical frameworks of diversity ranking and evaluation [2][7]. There are a growing amount of works in text summarization area

[6][9][12][20] which try to integrate diversity as part of the ranking function's properties. For example, Zhu et al. [20] proposes a unified ranking algorithm called GRASSHOPPER which is based on random walks over an absorbing Markov chain. Their method works by iteratively transforming the top-ranking nodes into absorbing nodes, effectively reducing the 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, thus ensuring that the novel nodes will be ranked higher. Li et al [12] approaches the diversity ranking from the optimization under constraints perspective. They propose a supervised method based on structural learning which incorporates diversity as a set of subtopic constraints. Then, they train a summarization model and enforce diversity through the optimization problem.

Our method differs from other diversity ranking methods in the rank propagation. Since most graph ranking models [6] [18][20] are inspired by the PageRank algorithm [4], they rely on eigenvector centrality to measure the importance of nodes. In contrast, our method uses the negative edges to propagate negative endorsements among nodes. High redundant nodes are those which receive a substantial amount of disapproval votes. Furthermore, the applications of negative edges in ranking model have been explored in related domains, such as trust ranking [8], social network mining [11][12], and complex question answering [1]. On the other hand, our method considers the negative edges to represent redundancy relationship among answers.

3. The Proposed Method

Our method focuses on two key aspects to find a diverse set of answer. First, a set of answer should contain many relevant facts pertaining to an information need. Second, Each answer should be novel or contain a distinct fact with respect to the other answers. To achieve that, we employ a graphical model to rank the relevant answers according to the balance between relevance and novelty. Two set of relations are represented by two signed graphs. First, the relevance relation is denoted by the positive edges between the question node and the answer nodes. On the other hand, the redundancy relation is represented by the negative edges between answer nodes. The negative sign is used to propagate a negative endorsement or disapproval vote between the nodes. The absolute value of edge weight represents the degree of similarity between answers. A formal description of our method is described below.

Given a question q and a set of n relevant answers, we first define $G = (V,E)$ as an undirected graph where V is a set of vertices representing the relevant answers, E is a set of edges representing the similarity between vertices where $E \subset V \times V$. Next, we can represent G as an n x n weighted matrix S where S_{ij} is a similarity score sim(i,j) of node i and j and sim(i,j) has a non-negative value. If i and j are unrelated, then $S_{ij} = 0$. Given S, we can derive an n x n normalized adjacency 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 $\sum_{j=1}^{n} A_{ij} = 1$. Next, given the question q, we define a vector r where each element r_i is the relevance score $rel(i,q)$ of node i given q. Then, we transform r into an n x n

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 $\sum_{j=1}^{n} B_{ij} = 1$. Given two probability distributions, we define the transition matrix P as:

$$P = dA + (1-d)B \tag{1}$$

where d is a damping factor with a real value from [0,1], A is the initial answer adjacency matrix, B is the question-answer 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 $\vec{\pi}P = \vec{\pi}$. At this stage, each answer can then be ranked according to its stationary distribution. At this point, we have derived a random walk model over the answer graph which incorporates both answer relevance and answer similarity. However, it does not take into account factual redundancy between answers.

In order to employ the negative edges to reduce factual redundancy, we modify the aforementioned 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 relevant answer vertices, E^{-} is a set of negative edges where $E^{-} \subset V \times V$. Then, an adjacency matrix M, corresponding to edge weights in G^{-}, is defined as an all-negative matrix of S where $M_{ij} = -\frac{S_{ij}}{\sum_{j=1}^{n} S_{ij}}$ and $\sum_{j=1}^{n} M_{ij} = -1$.

Next, similar to the process of deriving the transition matrix P, we define a modified transition matrix Q to incorporate the negative adjacency weights defined in M and the answer relevance defined in B. To ensure that Q is still 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. That is, $\sum_{j=1}^{n} Q_{ij} = 1$. Since all rows of M sum to -1 and all rows of B add up to 1, c is a function of d where $c = \frac{1+d}{1-d}$.

$$Q = dM + (1-d)cB \tag{2}$$

where M is an all-negative answer adjacency matrix. Since ergodicity properties still hold, Q has a unique stationary distribution $\vec{\pi}Q = \vec{\pi}$. Finally, we rank each node i according to its stationary probability $\vec{\pi}_i$. From the matrix notations, the simplified equation of the DiverseRank score can be formulated as follow:

$$DR^{t}(i) = (1-d)\cdot c\cdot \frac{rel(i,q)}{\sum_{i=1}^{n} rel(i,q)} - d \sum_{j \in adj(i)} \frac{sim(i,j)}{\sum_{k=1}^{n} sim(j,k)} DR^{t-1}(j) \tag{3}$$

Where d is a damping factor with a real value in [0,1] range. Additionally, d serves as a penalty factor of redundancy. c is a scaling factor defined as a function of d where $c = \frac{1+d}{1-d}$. $rel(i,q)$ is the relevance score of answer i given the question q. And

sim(i,j) is a similarity score of answer i and j. To estimate the value of $rel(i,q)$, we employ a sentence weighting function described in Allen et al. [3] as it is shown to consistently outperform other relevance models at the sentence level.

4. Experimental Evaluation

4.1 Data Sets

Two question answering data sets are used in the evaluation: a subset of Yahoo! Answers data set (YahooQA) used in Liu et al.'s work [15] and a complex interactive question answering test set (ciQA) used in TREC 2006 [10]. YahooQA data comprises subjective and ill-defined information needs formulated by the community members. The subjects of interests span widely from mathematics, general health, to wrestling. In contrast, ciQA data largely focus on the complex entity-relationship questions. Their information needs reflect those posed by intelligence analysts. From data quality perspective, YahooQA data are much noisier than ciQA data as they contain mostly informal linguistic expressions. To prepare YahooQA data set, we randomly select 100 questions and 10,546 answers from the top 20 most frequent categories (measured in terms of a number of responded answers) to use as a test set. A set of information nuggets for YahooQA is automatically created by matching relevant answers with the corresponding questions. The best answer chosen by askers for each question is marked as a vital nugget while the other answers are marked as an okay nugget. In the case of ciQA data set, 30 question topics and their free-form description are prepared by human assessors at National Institute of Standards and Technology (NIST).

4.2 Evaluation Settings

We employ the nugget pyramid metric to evaluate the diversity of the answer set. Generally, we assume that the factual diversity of the extracted answers can be measured in terms of a number of information nuggets the extracted sets have in common with the benchmark nuggets. The formulas to compute the pyramid F-score are described in [14]. In summary, the pyramid F-score is computed as a weighted harmonic mean (F-score) 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 extracted answer set. Following the standard procedure in TREC 2006, we set the evaluation parameters to $\beta = 3$ and l = 7,000 and use Pourpre [14] script version 1.1c to automatically compute the scores. Additionally, we further perform a fine-grained analysis of the algorithmic performance at varying sizes of answer set using a recall-by-length performance curve [13]. Better methods will produce curves that rise faster as more relevant and novel facts are included in the answer set.

Starting from the preprocessing step, we extract word features from a collection of answers by splitting answers into unigram tokens, removing non content-bearing

words, e.g., articles, conjunctions, prepositions, etc., and stemming the tokens using Porter Stemmer. After the preprocessing step, we use a vector-space model to retrieve the relevant answers. Free-form narrative field associated with each question is used as a query. The relevance scores between the answers and query are computed from the TFIFS weighting function [3]. The next step is to rerank the list of relevant answers obtained from the retrieval step. To achieve that, we first transform the list of relevant answers into an undirected graph with negative edges. Different edge weighting schemes based on inter-sentence similarity measures are employed. Next, the relevance scores of the retrieved answers and a query are calculated using various relevance models.

We compare the effectiveness of four baseline methods and the DiverseRank variants in answer re-ranking task. They are: Maximal marginal relevance (MMR) [5], SumBasic [17], Topic-Sensitive LexRank [18], and a backward ranking of Topic-Sensitive LexRank (henceforth LexRankInv). The last baseline method is defined to test whether diversity can be promoted by simply reversing the eigenvector centrality ranking. Next, several DiverseRank variants are tested based on the combinations of the initial ranking distribution and inter-sentence similarity functions. For the initial ranking distribution, we consider SumBasic (SB), TFISF-based relevance function (REL) described in section 3, a normalized sum of inverse document frequency (IDF) of the matching terms between the question and answer, and a uniform distribution $1/n$ (in the case of *novelty-only* variant). Next, two inter-sentence similarity measures are evaluated: N-Gram phrasal overlap (NG) and TFISF-weighted cosine similarity (TFISF). Moreover, we also evaluate two novelty-only variants of DiverseRank: 1+NG and 1+TFISF. The uniform distribution $1/n$ is assigned as the initial ranking distribution of the two novelty-only variants. After the ranking scores are calculated, we select top-k answers into the answer set. The default cut-off level for answer length is 7,000 characters.

5. Results and Discussion

5.1 Pyramid F-Scores

Table 1: The average F-Scores of the baseline and DiverseRank methods. The best methods are in bold.

Method	YahooQA		ciQA	
	F-Score	% Improvement	F-Score	% Improvement
MMR	0.2946	+14.52%	0.2486	+48.30%
SumBasic	0.2895	+16.54%	0.2956	+24.71%
LexRank	0.3163	+6.65%	0.3590	+2.69%
LexRankInv	0.2391	+41.12%	0.3516	+4.82%
DiverseRank	**0.3374**	-	**0.3686**	-

Table 1 shows the average pyramid F-scores of the baseline methods and the best DiverseRank variants. In both data sets, the proposed method significantly

outperforms all baseline methods, p-value<0.05. Considering the performance between random-walk based methods (LexRank vs. DiverseRank), DiverseRank also outperforms LexRank in both data sets although the improvements are relatively minor (6.65% and 2.69%), compared to those of other baselines. Furthermore, LexRankInv produces inferior scores to DiverseRank in both data sets.

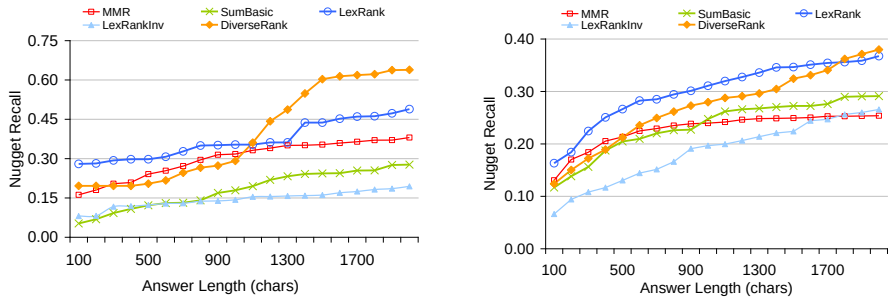
Table 2: The average F-Scores of the DiverseRank variants. The best methods are in bold.

Method	YahooQA	ciQA
SB+NG	0.2786	0.3433
SB+TFISF	0.2725	0.3454
IDF+NG	**0.3374**	0.3442
IDF+TFISF	0.2674	0.3445
REL+NG	0.3076	0.3400
REL+TFISF	0.2501	**0.3686**
1+NG	0.2348	0.3456
1+TFISF	0.2740	0.3439

Table 2 displays the performance of DiverseRank variants according to the combinations of relevance models and answer similarity measures. The best methods are IDF+NG and REL+TFISF, for YahooQA and ciQA respectively. As it can be seen, the best method differs between the two data sets. This is explained by the distinct characteristics of YahooQA and ciQA data. Questions and answers in ciQA originate from a more controlled environment since they are created by information professionals. Thus, they are more linguistically well-formed than those of YahooQA, which are created by the community members of unknown background. As such, ciQA data are generally less sparse and less noisy than YahooQA data, making cosine similarity-based variant (i.e., REL+TFISF) to perform better in ciQA case.

5.2 Recall-by-Length Performance Curves

Figure 2 shows recall-by-length curves of the baselines and the proposed method in each data set. In YahooQA case (figure 2a), DiverseRank starts to produce a significantly better performance than the best baseline method (LexRank) at the answer length of 1,200 characters. On the other hand, DiverseRank does not perform quite well in ciQA case (figure 2b). In this case, DiverseRank starts to outperform LexRank after the answer length of 1,800 characters. This outcome is not entirely unexpected. As a smaller answer set contains fewer number of information nuggets, therefore there are fewer items to be diversified. As the answer set continues to grow, DiverseRank continues to gain a better performance. Note that our results confirm the previous results in [10] in which a method that produces the best F-score at a predefined answer length does not necessarily perform equally effective across all incremental lengths.



2a. YahooQA **2b. ciQA**

Fig. 2. The recall-by-length performance curves of the best DiverseRank and baseline methods on YahooQA (A) and ciQA (B) data sets.

6. Conclusion

We propose a graph ranking model to find a diverse set of answers based on random walks over a negative-edge graph. Our main contribution is in the utilization of a graphical model to reduce redundancy within the answer set. First, given a complex question and a set of relevant answers, we represent the relevant answers as an answer graph whose nodes correspond to answers and edges correspond to the similarity between answers. Then, we assign a negative sign to edge weights in the answer graph to model the redundancy relationship between nodes. Then, the final ranking score for each answer node is derived from its long-term negative endorsement. The evaluation results show that our method outperforms most baseline methods. The analysis of recall-by-length performance curves suggests that the best baseline method performs better than DiverseRank at smaller answer lengths. This is explained by the fact that a smaller set of answers contains fewer facts, thus there is less room to promote diversity. As the size of answer set increases, DiverseRank eventually outperforms the baseline method.

Acknowledgments. This research work is supported in part from the NSF Career grant IIS 0448023, NSF CCF 0905291, NSF IIP 0934197, NSFC 90920005 and the Program of Introducing Talents of Discipline to Universities B07042 (China).

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. Agrawal, R., Gollapaudi, S., Halverson, A., and Ieong, S. (2009) Diversifying Search Results. In Proc. of WSDM'09, 5-14.
3. 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.

4. Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7).
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. Kelly, D. and Lin, J. (2007) Overview of the TREC 2006 ciQA Task. SIGIR Forum 41, 1 (June 2007), 107-116.
11. 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.
12. 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.
13. Lin, J. (2007) Is Question Answering Better Than Information Retrieval? Toward a Task-Based Evaluation Framework for Question Series. In Proc. of NAACL HLT 2007, Rochester, NY, 212-219.
14. Lin, J., and D., Demner-Fushman (2005) Automatically Evaluating Answers to Definition Questions. In Proc. of HLT/EMNLP, Vancouver, 931-938.
15. 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.
16. MacKinnon, I. and Vechtomova, O. (2008) Improving Complex Interactive Question Answering with Wikipedia Anchor Text. In Proc. of ECIR 2008, 438-445.
17. Nenkova, A. and Vanderwende, L. (2005) The impact of frequency on summarization. MSR-TR-2005-101.
18. 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.
19. 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.
20. 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.