Ranking Biomedical Text Passages for Relevance and Diversity
University of Wisconsin, Madison at TREC Genomics 2006

Andrew B. Goldberg, David Andrzejewski, Jurgen Van Gael, Burr Settles, Xiaojin Zhu, Mark Craven*

Department of Computer Sciences, *Department of Biostatistics & Medical Informatics, University of Wisconsin, Madison, WI 53705

Introduction

The 2006 Genomics track asks participants to design an information retrieval system that returns a diverse set of short text passages in response to a scientific query, such as *What is the role of PrnP in mad cow disease?* The goal is to select passages that highlight as many specific aspects of the question as possible. Systems are evaluated by document, passage, and aspect level metrics. We used off-the-shelf IR components and focused on query generation and reranking the query results. We focus here on one of our approaches to the latter task: a novel graph-theoretic ranking algorithm based on absorbing random walks, which aims to encourage both relevance and diversity among the highly ranked passages.

System Overview

Phase I: Indexing

- Split all documents into paragraphs / legal spans
- Index new files using an off-the-shelf IR engine (Lemur)

Phase II: Query Generation

- Automatically identify noun phrases (NPs) in topic description using an in-domain syntactic parser

  What is **the role** of **PrnP** in **mad cow disease**?

- Expand NPs with synonyms found in online resources

  PrnP ⇨ prion protein, GSS protein
  mad cow disease ⇨ Bovine Spongiform Encephalopathy

- Build structured query with criteria specifying that at least one synonym of each NP must be found

Phase III: Retrieval

- Execute query using IR engine
- Narrow down returned paragraphs into passages containing only relevant sentences

> In December 1984 a UK farmer called a veterinary surgeon to look at a cow that was behaving unusually. Seven weeks later the cow died. Early in 1985 more cows from the same herd developed similar clinical signs. In November 1986 **bovine spongiform encephalitis** (**BSE**) was first identified as a new disease, later reported in the veterinary press as a novel progressive **spongiform encephalopathy**. Later still the causal agent of **BSE** was recognized as an abnormal **prion protein**. Since the outset the story of **BSE** has been beset by problems.

Phase IV: Reranking

- Rerank the passages for relevance *and* diversity using:
  a) Lemur (baseline without reranking)
  b) Clustering
  c) Absorbing random walks

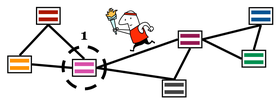
Reranking using Absorbing Markov Chain Random Walks

Key idea: Rank a set of passages such that

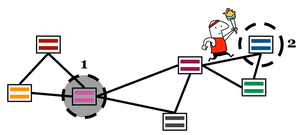
1. A highly ranked passage is representative of a local group in the set, i.e., it is similar to many other items. Ideally, these groups correspond to different aspects.
2. The top ranked passages cover many distinct groups.
3. An initial ranking is incorporated as prior knowledge.

We achieve these goals in a unified framework of *absorbing Markov chain random walks*.

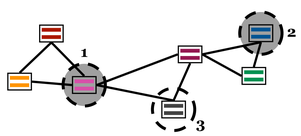
Example:



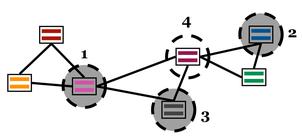
a) First passage chosen based on the stationary distribution of a random walk over passages in a graph.



b) First passage turns into an absorbing state. Second passage chosen based on expected number of visits before absorption.



c) The process repeats. Absorbing states "drag down" the importance of similar unranked passages to promote diversity.



d) Due to teleporting, a passage similar to previously ranked passages may be ranked highly if it had a high initial rank.

Complete Algorithm

Input: *W:* n x n weight matrix based on item similarity
r: prior distribution based on an initial ranking
$\lambda \in [0,1]$ trade-off parameter between W and r

Initialization:

1. Create transition matrix by normalizing rows of W

$$\tilde{P}_{ij} = \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}}$$

2. Interpolate each row with prior r to get teleporting walk

$$P = \lambda \tilde{P} + (1 - \lambda)\mathbf{1}\mathbf{r}^{\top}$$

Selecting the first item to rank:

1. Compute the unique stationary distribution of the walk

$$\pi = P^{\top}\pi$$

2. First item is state with largest stationary probability

$$g_1 = \arg\max_{i=1}^{n} \pi_i$$

Loop until all items are ranked:

1. Turn each ranked item $g \in G$ into an absorbing state

$$P_{gg} = 1, P_{gi} = 0, \forall i \neq g$$

$$P = \begin{bmatrix} \mathbf{I}_G & \mathbf{0} \\ R & Q \end{bmatrix}$$

2. Compute expected number of visits per unranked item

$$N = (\mathbf{I} - Q)^{-1}$$

$$\mathbf{v} = \frac{N^{\top}\mathbf{1}}{n - |G|}$$

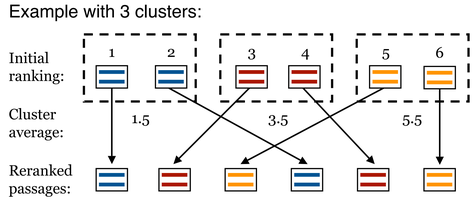
3. Select item with the maximum expected number of visits

$$g_{|G|+1} = \arg\max_{i=|G|+1}^{n} v_i$$

Reranking using Clustering

- Cluster passages using bag-of-words vectors and cosine similarity
- Assume clusters represent aspects
- Interleave results from clusters to achieve aspect diversity

Example with 3 clusters:



Results

Run	Document	Passage	Aspect
Lemur Ranking	0.2368	0.0188	0.1516
Clustering Reranking	0.2030	0.0137	0.1319
Absorbing Random Walk Reranking	0.2208	0.0159	0.1411

Document, passage, and aspect level mean average precision scores for the group's three automatic submissions.

Discussion

Document and passage level results:

- Mediocre results likely due to inadequate query generation.
- Poor topic parsing and no expansion led to returning zero passages for some of the 28 topics.
- Exact phrase matching query operators probably too restrictive.
- Correcting these issues and using additional resources (e.g., Gene Ontology, Unified Medical Language System Metathesaurus) might improve the results.
- Exploiting document-level features might help reveal additional relevant passages.

Aspect level results:

- Reranking did not improve aspect level results over baseline.
- Both reranking methods are sensitive to irrelevant documents, which appear diverse compared to relevant results, and thus end up incorrectly ranked high in the list.
- We assume similarity graph correlates with aspect similarity. If it does not, then the wrong kind of diversity is encouraged.
- To improve similarity, could use TF-IDF vectors, with IDF based on the current set of passages only.
- Could also use language-model KL-divergence as similarity.

Acknowledgments

AG was supported by a UW-Madison Graduate School Fellowship. DA was supported by an NLM training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM 5T15LM007359). BS and MC were supported in part by NSF grant IIS-0093016.

For Further Information

Please contact *goldberg@cs.wisc.edu*. More information on this work can be found at *www.cs.wisc.edu/~goldberg/trec.html*