Passage Relevance Models for Genomics Search

Jay Urbain
Elect Eng & Comp Sci
Milwaukee School of Engineering
Milwaukee, WI
urbain@msoe.edu

Ophir Frieder
Information Retrieval Lab
Illinois Institute of Technology
Chicago, IL
frieder@iit.edu

Nazli Goharian
Information Retrieval Lab
Illinois Institute of Technology
Chicago, IL
goharian@iit.edu

ABSTRACT

We present a passage relevance model for integrating semantic and statistical evidence of biomedical concepts and topics in context using the framework of a probabilistic graphical model. Component models of topics, concepts, terms, and document are represented as potential functions within a Markov Random Field, and the probability of a passage being relevant to a biologist's information need is represented as the joint distribution across all potential functions. Relevance model feedback of top ranked passages is used to improve distributional estimates of concepts and topics in context, and a dimensional indexing strategy is used for efficient aggregation of concept and term statistics. By integrating multiple sources of evidence including dependencies between topics, concepts, and terms, we seek to improve genomics literature passage retrieval precision. Using this model, we demonstrate statistically significant improvements in retrieval precision using a large genomics literature corpus.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Storage and Retrieval

General Terms: Algorithms, Experimentation, Theory

Keywords: Information retrieval, genomics search, semantic search, relevance modeling, passage retrieval, dimensional modeling, graphical models, passage retrieval.

1. PASSAGE RELEVANCE MODEL

As shown in Figure 1, our proposed *passage relevance model* represents the joint probability of query Q and passage P as an undirected graphical model. Edges in the graph define conditional independence assumptions between component *topic, concept, term*, and *document* $\theta_p, \theta_c, \theta_t, \theta_d$ models, respectively. Random variable P represents the distribution of features present in the passage without relevance estimates, and P_R represents a refinement to this distribution using a *relevant* set of passages. We use the top ranked passages retrieved from the model without using the relevant set as an estimate for P_R.

Based on conditional independence assumptions, the model is factorized into a set of maximal cliques. The joint probability distribution is written as a product of potential functions $\psi(c)$ over the maximal cliques in the graph (1).

$$p(Q,P) = \frac{1}{Z}\prod_{c\in C(G)}\psi(c) \tag{1}$$

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

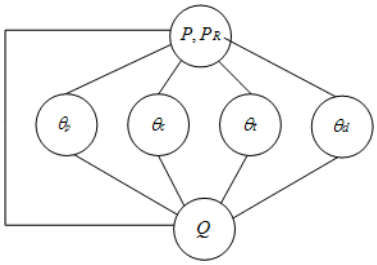

Figure 1 Passage Relevance Model

Potential functions in the passage retrieval model are defined for *topic, concept, term*, and *document* cliques as:

$$\psi(Q,P,P_R,\theta_p) \quad \psi(Q,P,P_R,\theta_c) \quad \psi(Q,P,P_R,\theta_t) \quad \psi(Q,P,\theta_d) \tag{2}$$

The joint probability of a query Q and passage P across all potential functions results in the following:

$$p(Q,P) \propto \sum_{c\in C(G)}\log(\psi(c)) \tag{3}$$

2. TOPIC RELEVANCE MODEL

The objective of modeling *topic relevance* is to directly address the issue of learning a topic model that is *relevant* to the latent structure, or *topic*, of a user query by capturing the probability of each term over all other terms in a relevant set of passages. We construct a model estimating the relevance of any given passage to a *query* topic from the probability of relevance of its component terms (4,5,6).

$$p(w_i \mid R) = \frac{p(R \mid w_i)\,p(w_i)}{p(R)} \tag{4}$$

$$p(w_i) = \frac{p(w \mid R)}{p(w \mid \neg R)} = \frac{C^{|S_R|}_{w_i\in S_R} + \beta}{C^{|S|}_{w_i\in S, w_i\notin S_R} + |S_R|\cdot\beta} \tag{5}$$

$|S_R|$ is the length of the set of relevant passages.

$C^{|S_R|}_{w_i\in S_R}$ is the count of relevant passages containing w_i.

$C^{|S|}_{w_i\in S, w_i\notin S_R}$ count of all paragraphs containing of w_i.

β is a smoothing parameter set to zero.

$$p(R \mid w_i) \approx \alpha\left(\frac{1}{Z}\sum_{j\neq i}^{K} p(R \mid w_k)\right) + (\alpha - 1)\cdot p(w_i) \tag{6}$$

We define the probability of a query q for a given passage m_j using the sum rule from the underlying *topic relevance* model θ_R (7).

$$p(q \mid m_j) = p(m_j \mid \theta_R) = \sum_i^W p(R \mid w_i) \tag{7}$$

As a proxy for the relevant set of passages, we sample terms from the top *30* ranked passages containing at least one resolved *concept* using the *passage retrieval model* (Figure 1) *without*

using the discrete random variable representing *topic relevance* which we seek here to create. The full model is then evaluated on the top *500* retrieved passages. In Table 1, we show the top 30 *topic relevance* terms learned for queries 200, 201, and 202 of the 2007 TREC Genomics track. Qualitatively, the relevance terms learned for each query appear highly relevant. It is especially interesting to note that many of the top *topic relevance* terms were not present in the query, and were not identified as term variants by our normalization procedure or as concept synonyms from external knowledge sources.

Table 1 Topic Relevance

Query 200	Query 201	Query 202
BILAG (0.6010)	B-RAF (0.54793)	1gangliosid (0.5424)
lupu (0.5650)	mutat (0.5039)	brain (0.5010)
anticardiolipin (0.3960)	RAF (0.4834)	gangliosid (0.4949)
immunodiffus (0.3870)	melanoma (0.4536)	accumul (0.4146)
isle (0.3750)	activ (0.4403)	abnorm (0.4008)
system (0.3331)	mutation (0.3661)	diseas (0.2690)
erythematosu (0.2954)	cell (0.3649)	asialo (0.2393)
antibodi (0.2820)	ERK (0.2508)	neuron (0.2323)
index (0.2776)	gene (0.2132)	protein (0.2067)
diseas (0.2488)	RAS (0.19943)	patient (0.1990)
activ (0.24514)	pathwai (0.1804)	cell (0.1985)
measur (0.2432)	human (0.1781)	lysosom (0.1895)
anticoagul (0.2320)	cancer (0.16233)	respons (0.1836)
clinic (0.2193)	autoinhibit (0.1617)	human (0.1793)
bacon (0.1807)	express (0.1570)	promin (0.1724)
patient (0.1551)	growth (0.1447)	mice (0.16800)
EM (0.1492)	phosphoryl (0.1183)	clinic (0.1673)
ISI (0.1425)	focus (0.1162)	gangliosidosi (0.1647)
hay (0.1404)	signal (0.1152)	phenotyp (0.1599)
score (0.1291)	tumor (0.1148)	storag (0.1559)
SLE (0.1196)	RAF1 (0.1128)	apoptosi (0.1326)

200: What serum [proteins] change expression in association with high disease activity in lupus?
201: What [mutations] in the Raf gene are associated with cancer?
202: What [drugs] are associated with lysosomal abnormalities in the nervous system?
Note: Terms shown in stemmed form, acronyms have been capitalized.

3. CONCEPT, TERM & DOC MODELS

The likelihood of each sentence being generated for a given concept, and the likelihood of each sentence being generated for a given term is determined from the concept-word and term-word co-occurrence distributions respectively (Urbain, Frieder, Goharian, 2008). A Jelinek-Mercer language model is used to capture document level evidence.

4. RESULTS

Results on the 2007 TREC Genomics track of *162,000* full-text documents, and 36 topic queries (Hersh et al., 2007) are listed in Table 2. Topic modeling using relevance outperformed automatically generated topic models by *28.07%,* and is computationally efficient. The full retrieval model outperforms models of query terms, concepts, document, or passage relevance alone. The model exceeds the top results in each category of retrieval as assessed by the 2007 TREC Genomics track and the results are *statistically significant* for automatic document and passage retrieval. All results are for *automatic retrieval*.

Table 2 Results 2007 TREC Genomics collection (MAP)

Model	Doc	Passage	Passage2	Aspect
Top TREC*	0.3105	0.0976	0.1097	0.2494
Median TREC	0.1954	0.0565	0.0391	0.1272
TREC 2007 Submission	0.2385	0.09742	0.1647	0.05164
Document model	0.2363	-	-	-
Topic model No relevance	0.2034	-	-	-
Topic-relevance model	0.2605	0.0898	0.0452	0.1383
Concept model	0.3381	0.1087	0.0579	0.1907
Term model	0.3226	0.1053	0.0557	0.1856
Concept + Term models	0.3443	0.1100	0.0588	0.2145
Doc+Concept +Term+Topic -relevance *Min Spanning Passage*	0.3554 (+14.46%) *p=0.0582*	**0.1214 (+24.39%)** ***p=0.0321***†	0.0681 (-37.92%)	0.2412 (-3.29%)
Doc+Concept +Term+Topic -relevance *Max Spanning Passage*	**0.3576 (+15.17%)** ***p=0.0504***	0.1093 (+11.99%)	**0.1280 (+16.68%)** ***p=0.0834***	**0.2596 (+4.08%)**

†Statistically significant using Wilcoxon signed rank test ($p<0.05$).

REFERENCES

Hersh W., et al. (2007). TREC 2007 Genomics Track Overview. The Sixteenth Text REtrieval Conference Proceedings.

Steyvers, M. (2006). Probabilistic Topic Models. In Landauer, T., McNamara, D., Dennis, S., & Kintch W. (eds), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum.

Urbain, J., Goharian, N., & Frieder, O. (2007, November). IIT TREC 2007 Genomics Track: Using Concept-Based Semantics in Context for Genomics Literature Passage Retrieval. The Sixteenth Text REtrieval Conference (TREC 2007) Conference Proceedings.

Urbain, J., Frieder, O., & Goharian, N. (2008 to appear). Probabilistic Passage Models for Semantic Search of Genomics Literature. Journal of the American Society of Information Science.