

Bridging the Terminology Gap in Web Archive Search

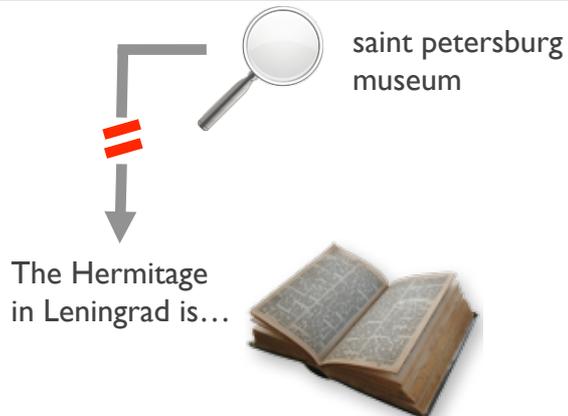
Klaus Berberich

Motivation

Web archives preserve documents that were published a long time ago. When searching these archives, today's users formulate keyword queries using today's terminology. Terminology, though, evolves constantly! There is thus an **ever-widening terminology gap** between the keyword queries formulated by today's users and the archived documents. As a consequence, **old documents** that are highly relevant to the user's information need are **often not retrieved!**

Our Approach

To bridge this terminology gap, we **reformulate the user's keyword query** to retrieve old but highly relevant documents.



Query Reformulation

Given a keyword query $q = \langle q_1, \dots, q_m \rangle$ formulated using today's terminology, we determine a reformulated query $q' = \langle q'_1, \dots, q'_m \rangle$ having the same length that paraphrases the user's information need using terminology prevalent at a target time in the past. Three desiderata for a good query reformulation are:

Similarity: Keywords in q' should have a high degree of semantic similarity with their respective counterpart in q .

Coherence: Keywords in q' should be frequently used together at the target time

Popularity: Keywords in q' should be in common use at the target time.

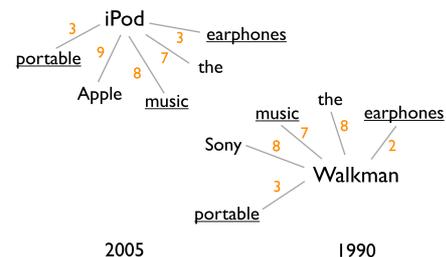
We propose a **Hidden-Markov Model (HMM)** taking into account these desiderata. The "goodness" of a query reformulation is defined as

$$P(q | q') = P(q_1) \cdot P(q_2 | q'_1) \cdot \prod_{i=2}^m P(q_i | q'_{i-1}) \cdot P(q_i | q'_i)$$

The best k query reformulations according to this model can be **determined efficiently** using a combination of the **Viterbi algorithm** and **A* search**.

Across-Time Semantic Similarity

We want to assess the degree of **semantic similarity** between two terms **when used at different times**. To this end, we leverage the **terms' contexts** as observed in documents published at the respective times.



From the fact that iPod and Walkman **co-occur frequently with similar terms** in 2005 and 1990, respectively, we infer a high degree of across-time semantic similarity.

Experimental Results

Experiments conducted on **New York Times Annotated Corpus** that contains 1.8M NYT articles published between 1987 and 2007.

	airbus a380 → 2000
1.	airbus industries
2.	a3xx superjumbo
3.	airbus superjumbo

	kyoto protocol → 1990
1.	berenter greenhouse
2.	greenhouse_effect warming
3.	greenhouse_effect gases

	tony blair prime minister → 1990
1.	margaret_thatcher prime minister
2.	yitzshak_shamir prime minister
3.	Vacek minister prime

References

M. Federico and N. Bertoldi
Statistical Cross-Language IR using N-Best Query Translations, SIGIR 2002
H. Rubenstein and J. B. Goodenough
Contextual correlates of synonymy, CACM 8(10), 1965.
N. Tahmasebi, T. Iofciu, T. Risse, C. Niederée, W. Siberski
Terminology Evolution in Web Archiving: Open Issues, IWAW 2008

Other PhD Projects

I. Arıkan, S. Bedathur, K. Berberich
Time will Tell: Leveraging Temporal Expressions in IR, WSDM 2009
K. Berberich, S. Bedathur, T. Neumann, G. Weikum
A Time Machine for Text Search, SIGIR 2007
K. Berberich, S. Bedathur, G. Weikum
A Pocket Guide to Web History, SPIRE 2007