Bridging the Terminology Gap in Web Archive Search

Klaus Berberich, Srikanta Bedathur,
Mauro Sozio, Gerhard Weikum

Max-Planck Institute for Informatics,
Saarbrücken, Germany

mpii

LiWA
Living Web Archives

- http://www.liwa-project.eu

- European Union FP7 project that develops next generation web archiving technologies

Web Archives

- Archived contents increasingly made available on the Web – Web content increasingly archived

http://archives.timesonline.co.uk
Issues since 1785 digitized

http://archive.org/web
150B web pages archived since 1996

- Web archives play an important role in providing access and preserving our cultural heritage

What is the Terminology Gap?

- Terminology evolves constantly! Consider, e.g.,
  - Saint Petersburg@2009 ≈ Leningrad@1978
  - Firefighter@2005 ≈ Fireman@1968
  - Person month@2000 ≈ Man month@1980
- Keyword search on web archives suffers from the terminology gap between today's queries and yesterday's documents

saint petersburg museum

2009


History preserved in shro...
The Times | September 20, 1978
An extraordinary exhibition, on loan from Leningrad's Hermitage Museum, is to be displayed this winter (Nov-

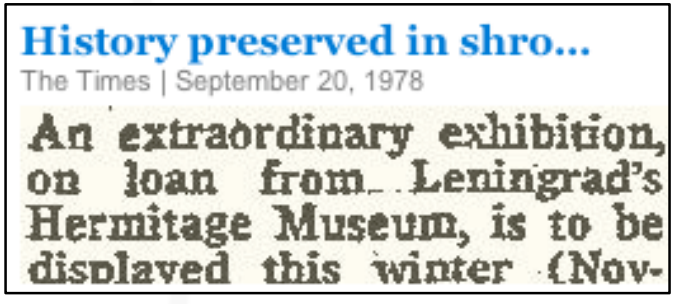
1978

Our Approach

- Reformulate keyword queries to also retrieve old but highly-relevant documents

saint petersburg museum → *leningrad hermitage* →

2009 **1978**


History preserved in shro...
The Times | September 20, 1978
An extraordinary exhibition, on loan from Leningrad's Hermitage Museum, is to be displayed this winter (Nov-

- Given a keyword query *q* formulated using terminology valid at a reference time *R*, we identify a query reformulation *q'* that paraphrases the same information need using terminology valid at a target time *T*

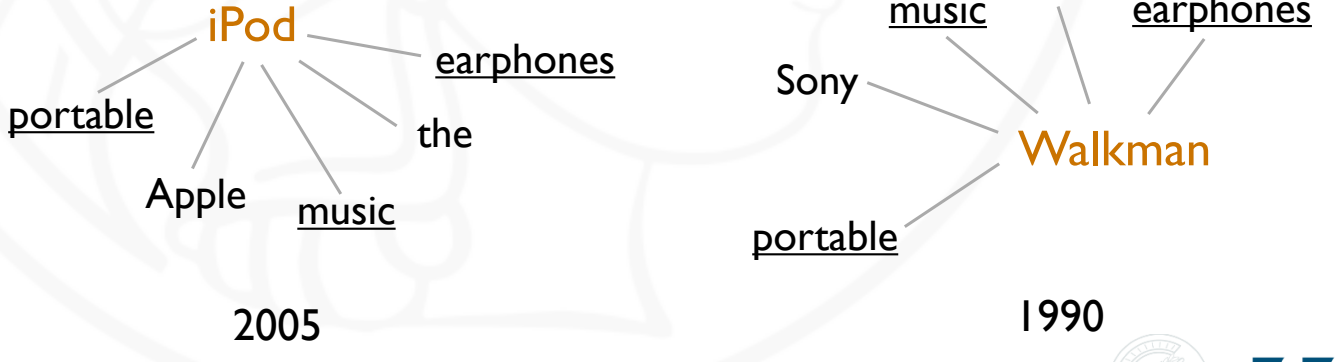
Outline

- Motivation
- Across-Time Semantic Similarity
- Query Reformulation
- Implementation Issues
- Experiments
- Conclusion & Future Work

Across-Time Semantic Similarity

- Quantify the degree of semantic similarity between two terms when used at different times
  - iPod@2005 ~ Walkman@1990
  - Saint Petersburg@2009 ~ Leningrad@1978

- Idea: Compare terms' contexts at the times (i.e., frequently co-occurring terms)



Across-Time Semantic Similarity

- Use term (co-)occurrence statistics computed on documents published during *T* and *R*

$$P(\, u@R \,|\, w@R \,) = \frac{cooc(w@R, v@R)}{\sum_{z \in \mathcal{V}} cooc(w@R, z@R)}$$

$$P(u@R) = \frac{freq(\, u@R \,)}{\sum_{z \in \mathcal{V}} freq(\, z@R \,)}$$

- Generative model according to which *v@T* has high probability of generating *u@R* if there is large overlap in their respective term contexts

$$P(\, u@R \,|\, v@T \,) = \sum_{w \in \mathcal{V}} P(\, u@R \,|\, w@R \,) \cdot P(\, w@T \,|\, v@T \,)$$

Outline

- Motivation
- Across-Time Semantic Similarity
- Query Reformulation
- Implementation Issues
- Experiments
- Conclusion & Future Work

Query Reformulation

- Problem: Given $q@R = \langle q_1,\ldots,q_m \rangle$ find a good query reformulation $q'@T = \langle q'_1,\ldots,q'_m \rangle$

  What makes up a good query reformulation?
  - Similarity, i.e., q_i and q_i' should have high a degree of across-time semantic similarity
  - Coherence, i.e., q'_i and q'_{i-1} should co-occur frequently at time T to avoid combining unrelated terms, e.g.,
    leningrad smithsonian@1978
  - Popularity, i.e., q'_i should occur frequently at time T to avoid unlikely query reformulations, e.g.,
    saarbruecken saarlandmuseum@1978

Query Reformulation

- Hidden Markov model (HMM) that considers these three desiderata
  - Similarity measured as $P(q_i @ R \mid q'_i @ T)$
  - Coherence measured as $P(q'_i @ T \mid q'_{i-1} @ T)$
  - Popularity measured as $P(q'_i @ T)$
- Good query reformulations correspond to state sequences that have a high probability of being traversed while generating our original query q

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

Query Reformulation

- Top-*k* query reformulations determined using a combination of Viterbi algorithm and A* Search
- Viterbi algorithm determines the best state sequence using dynamic programming
- A* Search identifies top-*k* query reformulations leveraging information memoized by Viterbi
- Time complexity in $O(m \cdot |V|^2)$
- Space complexity in $O(m \cdot |V|)$

m = query length
$|V|$ = overall number of terms

Outline

- Motivation
- Across-Time Semantic Similarity
- Query Reformulation
- Implementation Issues
- Experiments
- Conclusion & Future Work

Implementation Issues

- Safe state pruning, i.e., we ignore all terms *v@T* that have zero across-time semantic similarity with all query terms q_i
- Additional heuristic state pruning, i.e., for each q_i consider only the κ terms *v@T* having highest across-time semantic similarity
- Precomputation, i.e., we limit choices of *R* and *T* to calendar years and precompute values *P(u@T | v@T)* and *P(u@T)* accordingly

Outline

- Motivation
- Across-Time Semantic Similarity
- Query Reformulation
- Implementation Issues
- Experiments
- Conclusion & Future Work

Experimental Setup

- Dataset: New York Times Annotated Corpus containing 1.8M articles from 1987 – 2007
- Simple phrase extraction based on Wikipedia titles to capture multi-term expression (e.g., john_lennon, disk_operating_system, etc.)
- Implementation: Java, data kept in Oracle 10g DB

Experimental Results

- Across-time semantic similarity

u R/T	**pope_benedict** 2005 / 1990	**starbucks** 2005 / 1990	**linux** 2005 / 1990	**mp3** 2005 / 1990
1	alexander_pope	dunkin_donuts	unix_operating_system	audio_cd
2	the_pope	dunkin	unix_systems	digital_audio
3	cardinal_ratzinger	donuts	unix_international	computer_files
4	joseph_cardinal_ratzinger	coffee_shops	the_operating_system	s_files
5	pope_john_paul	cup_of_coffee	disk_operating_system	the_rockford_files
6	pope_john_paul_ii	a_cup_of_coffee	dos_operating_system	rockford_files
7	conservative_catholics	coffee_cup	operating_system	audio_systems
8	polish-born	coffee_shop	operating_systems	audio_tapes
9	irish_catholics	morning_coffee	os	audio_equipment
10	frantisek_cardinal_tomasek	coffee_filter	os_2	audio_clips

Experimental Results

- Across-time semantic similarity

u R/T	**pope_benedict** 2005 / 1990	**starbucks** 2005 / 1990	**linux** 2005 / 1990	**mp3** 2005 / 1990
1	alexander_pope	dunkin_donuts	unix_operating_system	audio_cd
2	the_pope	dunkin	unix_systems	digital_audio
3	cardinal_ratzinger	donuts	unix_international	computer_files
4	joseph_cardinal_ratzinger	coffee_shops	the_operating_system	s_files
5	pope_john_paul	cup_of_coffee	disk_operating_system	the_rockford_files
6	pope_john_paul_ii	a_cup_of_coffee	dos_operating_system	rockford_files
7	conservative_catholics	coffee_cup	operating_system	audio_systems
8	polish-born	coffee_shop	operating_systems	audio_tapes
9	irish_catholics	morning_coffee	os	audio_equipment
10	frantisek_cardinal_tomasek	coffee_filter	os_2	audio_clips

Experimental Results

- Across-time semantic similarity

u R/T	**pope_benedict** 2005 / 1990	**starbucks** 2005 / 1990	**linux** 2005 / 1990	**mp3** 2005 / 1990
1	alexander_pope	dunkin_donuts	unix_operating_system	audio_cd
2	the_pope	dunkin	unix_systems	digital_audio
3	cardinal_ratzinger	donuts	unix_international	computer_files
4	joseph_cardinal_ratzinger	coffee_shops	the_operating_system	s_files
5	pope_john_paul	cup_of_coffee	disk_operating_system	the_rockford_files
6	pope_john_paul_ii	a_cup_of_coffee	dos_operating_system	rockford_files
7	conservative_catholics	coffee_cup	operating_system	audio_systems
8	polish-born	coffee_shop	operating_systems	audio_tapes
9	irish_catholics	morning_coffee	os	audio_equipment
10	frantisek_cardinal_tomasek	coffee_filter	os_2	audio_clips

Experimental Results

- Across-time semantic similarity

u R/T	**pope_benedict** 2005 / 1990	**starbucks** 2005 / 1990	**linux** 2005 / 1990	**mp3** 2005 / 1990
1	alexander_pope	dunkin_donuts	unix_operating_system	audio_cd
2	the_pope	dunkin	unix_systems	digital_audio
3	cardinal_ratzinger	donuts	unix_international	computer_files
4	joseph_cardinal_ratzinger	coffee_shops	the_operating_system	s_files
5	pope_john_paul	cup_of_coffee	disk_operating_system	the_rockford_files
6	pope_john_paul_ii	a_cup_of_coffee	dos_operating_system	rockford_files
7	conservative_catholics	coffee_cup	operating_system	audio_systems
8	polish-born	coffee_shop	operating_systems	audio_tapes
9	irish_catholics	morning_coffee	os	audio_equipment
10	frantisek_cardinal_tomasek	coffee_filter	os_2	audio_clips

Experimental Results

- Across-time semantic similarity

u R/T	**pope_benedict** 2005 / 1990	**starbucks** 2005 / 1990	**linux** 2005 / 1990	**mp3** 2005 / 1990
1	alexander_pope	dunkin_donuts	unix_operating_system	audio_cd
2	the_pope	dunkin	unix_systems	digital_audio
3	cardinal_ratzinger	donuts	unix_international	computer_files
4	joseph_cardinal_ratzinger	coffee_shops	the_operating_system	s_files
5	pope_john_paul	cup_of_coffee	disk_operating_system	the_rockford_files
6	pope_john_paul_ii	a_cup_of_coffee	dos_operating_system	rockford_files
7	conservative_catholics	coffee_cup	operating_system	audio_systems
8	polish-born	coffee_shop	operating_systems	audio_tapes
9	irish_catholics	morning_coffee	os	audio_equipment
10	frantisek_cardinal_tomasek	coffee_filter	os_2	audio_clips

Experimental Results

- Across-time semantic similarity

u R/T	**pope_benedict** 2005 / 1990	**starbucks** 2005 / 1990	**linux** 2005 / 1990	**mp3** 2005 / 1990
1	alexander_pope	dunkin_donuts	unix_operating_system	audio_cd
2	the_pope	dunkin	unix_systems	digital_audio
3	cardinal_ratzinger	donuts	unix_international	computer_files
4	joseph_cardinal_ratzinger	coffee_shops	the_operating_system	s_files
5	pope_john_paul	cup_of_coffee	disk_operating_system	the_rockford_files
6	pope_john_paul_ii	a_cup_of_coffee	dos_operating_system	rockford_files
7	conservative_catholics	coffee_cup	operating_system	audio_systems
8	polish-born	coffee_shop	operating_systems	audio_tapes
9	irish_catholics	morning_coffee	os	audio_equipment
10	frantisek_cardinal_tomasek	coffee_filter	os_2	audio_clips

Experimental Results

- Across-time semantic similarity

u R/T	**pope_benedict** 2005 / 1990	**starbucks** 2005 / 1990	**linux** 2005 / 1990	**mp3** 2005 / 1990
1	alexander_pope	dunkin_donuts	unix_operating_system	audio_cd
2	the_pope	dunkin	unix_systems	digital_audio
3	cardinal_ratzinger	donuts	unix_international	computer_files
4	joseph_cardinal_ratzinger	coffee_shops	the_operating_system	s_files
5	pope_john_paul	cup_of_coffee	disk_operating_system	the_rockford_files
6	pope_john_paul_ii	a_cup_of_coffee	dos_operating_system	rockford_files
7	conservative_catholics	coffee_cup	operating_system	audio_systems
8	polish-born	coffee_shop	operating_systems	audio_tapes
9	irish_catholics	morning_coffee	os	audio_equipment
10	frantisek_cardinal_tomasek	coffee_filter	os_2	audio_clips

Experimental Results

- Query reformulations

q R/T	**george_bush speech** 2005 / 1990	**colin_powell iraq** 2005 / 1990	**kyoto protocol** 2005 / 1990
1	george_bush speech	james_baker saddam_hussein	berenter greenhouse
2	president_ronald_reagan excerpts	james_baker hussein	greenhouse_effect warming
3	barbara_bush commencement	james_baker iraq	greenhouse_effect gases
q R/T	**tony_blair prime minister** 2005 / 1990	**christo gates** 2005 / 1995	**nintendo ds** 2005 / 1990
1	margaret_thatcher prime minister	jeanne-claude christo	game_boy nintendo
2	yitshak_shamir prime minister	christo reichstag	video-game nintendo
3	vacek minister prime	christo the_reichstag	galoob nintentdo

Experimental Results

- Query reformulations

q R/T	**george_bush speech** 2005 / 1990	**colin_powell iraq** 2005 / 1990	**kyoto protocol** 2005 / 1990
1	george_bush speech	james_baker saddam_hussein	berenter greenhouse
2	president_ronald_reagan excerpts	james_baker hussein	greenhouse_effect warming
3	barbara_bush commencement	james_baker iraq	greenhouse_effect gases
q R/T	**tony_blair prime minister** 2005 / 1990	**christo gates** 2005 / 1995	**nintendo ds** 2005 / 1990
1	margaret_thatcher prime minister	jeanne-claude christo	game_boy nintendo
2	yitshak_shamir prime minister	christo reichstag	video-game nintendo
3	vacek minister prime	christo the_reichstag	galoob nintentdo

Experimental Results

- Query reformulations

q R/T	**george_bush speech** 2005 / 1990	**colin_powell iraq** 2005 / 1990	**kyoto protocol** 2005 / 1990
1	george_bush speech	james_baker saddam_hussein	berenter greenhouse
2	president_ronald_reagan excerpts	james_baker hussein	greenhouse_effect warming
3	barbara_bush commencement	james_baker iraq	greenhouse_effect gases
q R/T	**tony_blair prime minister** 2005 / 1990	**christo gates** 2005 / 1995	**nintendo ds** 2005 / 1990
1	margaret_thatcher prime minister	jeanne-claude christo	game_boy nintendo
2	yitshak_shamir prime minister	christo reichstag	video-game nintendo
3	vacek minister prime	christo the_reichstag	galoob nintentdo

Experimental Results

- Query reformulations

q R/T	**george_bush speech** 2005 / 1990	**colin_powell iraq** 2005 / 1990	**kyoto protocol** 2005 / 1990
1	george_bush speech	james_baker saddam_hussein	berenter greenhouse
2	president_ronald_reagan excerpts	james_baker hussein	greenhouse_effect warming
3	barbara_bush commencement	james_baker iraq	greenhouse_effect gases
q R/T	**tony_blair prime minister** 2005 / 1990	**christo gates** 2005 / 1995	**nintendo ds** 2005 / 1990
1	margaret_thatcher prime minister	jeanne-claude christo	game_boy nintendo
2	yitshak_shamir prime minister	christo reichstag	video-game nintendo
3	vacek minister prime	christo the_reichstag	galoob nintentdo

Experimental Results

- Query reformulations

q R/T	**george_bush speech** 2005 / 1990	**colin_powell iraq** 2005 / 1990	**kyoto protocol** 2005 / 1990
1	george_bush speech	james_baker saddam_hussein	berenter greenhouse
2	president_ronald_reagan excerpts	james_baker hussein	greenhouse_effect warming
3	barbara_bush commencement	james_baker iraq	greenhouse_effect gases
q R/T	**tony_blair prime minister** 2005 / 1990	**christo gates** 2005 / 1995	**nintendo ds** 2005 / 1990
1	margaret_thatcher prime minister	jeanne-claude christo	game_boy nintendo
2	yitshak_shamir prime minister	christo reichstag	video-game nintendo
3	vacek minister prime	christo the_reichstag	galoob nintentdo

Experimental Results

- Query reformulations

q R/T	**george_bush speech** 2005 / 1990	**colin_powell iraq** 2005 / 1990	**kyoto protocol** 2005 / 1990
1	george_bush speech	james_baker saddam_hussein	berenter greenhouse
2	president_ronald_reagan excerpts	james_baker hussein	greenhouse_effect warming
3	barbara_bush commencement	james_baker iraq	greenhouse_effect gases
q R/T	**tony_blair prime minister** 2005 / 1990	**christo gates** 2005 / 1995	**nintendo ds** 2005 / 1990
1	margaret_thatcher prime minister	jeanne-claude christo	game_boy nintendo
2	yitshak_shamir prime minister	christo reichstag	video-game nintendo
3	vacek minister prime	christo the_reichstag	galoob nintentdo

Experimental Results

- Query reformulations

q R/T	**george_bush speech** 2005 / 1990	**colin_powell iraq** 2005 / 1990	**kyoto protocol** 2005 / 1990
1	george_bush speech	james_baker saddam_hussein	berenter greenhouse
2	president_ronald_reagan excerpts	james_baker hussein	greenhouse_effect warming
3	barbara_bush commencement	james_baker iraq	greenhouse_effect gases
q R/T	**tony_blair prime minister** 2005 / 1990	**christo gates** 2005 / 1995	**nintendo ds** 2005 / 1990
1	margaret_thatcher prime minister	jeanne-claude christo	game_boy nintendo
2	yitshak_shamir prime minister	christo reichstag	video-game nintendo
3	vacek minister prime	christo the_reichstag	galoob nintentdo

Outline

- Motivation
- Across-Time Semantic Similarity
- Query Reformulation
- Implementation Issues
- Experiments
- Conclusion & Future Work

Conclusion

- Terminology evolution is an important issue that needs to be addressed in web archives
- Novel measure of across-time semantic similarity
- Query reformulation approach based on a HMM
- Promising initial experimental results

Future Work

- Refine the model to deal with multi-term expressions in a more principled manner
- Further improve the efficiency of best-*k* query reformulation computation
- Overcome restricted choice of *R* and *T*

Thanks!

Questions & Ideas?