Wikipedia Pages as Entry Points for Book Search

Marijn Koolen[1,2] Gabriella Kazai[1] Nick Craswell[1]
[1] Microsoft Research, Cambridge, UK
[2] University of Amsterdam, the Netherlands

WSDM'09
Barcelona, 10 February, 2009

Outline

- **Introduction**
- Wikipedia Coverage
  - ⋆ Wikipedia coverage of search topics
  - ⋆ Wikipedia coverage of book topics
- Wikipedia as intermediary
  - ⋆ Query Expansion
  - ⋆ Topical Closeness
- Experiments & results
- Conclusion

Introduction

- Thanks to mass book digitisation efforts, large book collections now available online
- Books online thanks to mass-digitisation projects
  - ⋆ Million Books Project, Google Book Search

Introduction

- Thanks to mass book digitisation efforts, large book collections now available online
- Books online thanks to mass-digitisation projects
  - ⋆ Million Books Project, Google Book Search
- Significant repository of (untapped) knowledge:

  "*For hundreds of years books have been the repositories for the worlds most trusted, authoritative knowledge*. " – Cliff Guren, Live Search
- Access through book search (content or metadata)

Our Approach

- Use Wikipedia as an intermediary between users and books:
  - ⋆ Our knowledge of the world is (for a significant part) stored in books
  - ⋆ Encyclopedias summarise this world knowledge:

    "...*the purpose of an encyclopedia is to collect knowledge disseminated around the globe*..." – Diderot

Our Approach

- Use Wikipedia as an intermediary between users and books:
  - ⋆ Our knowledge of the world is (for a significant part) stored in books
  - ⋆ Encyclopedias summarise this world knowledge:

    "...*the purpose of an encyclopedia is to collect knowledge disseminated around the globe...*" – Diderot

Research Questions

- Many search topics have an entry in Wikipedia:
  - ⋆ Can we automatically extract useful search terms from related Wikipedia pages to improve retrieval effectiveness of a book search system?

Research Questions

- Many search topics have an entry in Wikipedia:
  - ⋆ Can we automatically extract useful search terms from related Wikipedia pages to improve retrieval effectiveness of a book search system?
- Many book topics have corresponding Wiki pages as well
- Wikipedia has many links between related topics:

Research Questions

- Many search topics have an entry in Wikipedia:
  - ⋆ Can we automatically extract useful search terms from related Wikipedia pages to improve retrieval effectiveness of a book search system?
- Many book topics have corresponding Wiki pages as well
- Wikipedia has many links between related topics:
  - ⋆ Is the link distance between search topics and book topics in Wikipedia related to relevance and can we use this to improve retrieval effectiveness?

Wikipedia Coverage

- Our approach relies on two assumptions:

1. Wikipedia covers many user search topics
2. Wikipedia covers the topics found in books

Wikipedia Coverage

- Our approach relies on two assumptions:
  1. Wikipedia covers many user search topics
  2. Wikipedia covers the topics found in books
- Two intuitions support these assumptions:
  1. Wikipedia is collectively written, on topics of interest
  2. Encyclopedias collect and summarise human knowledge
- Do we have more than just intuitions?

Wikipedia Coverage of Search Topics

- Do Wikipedia entries cover topics searched for by web users?

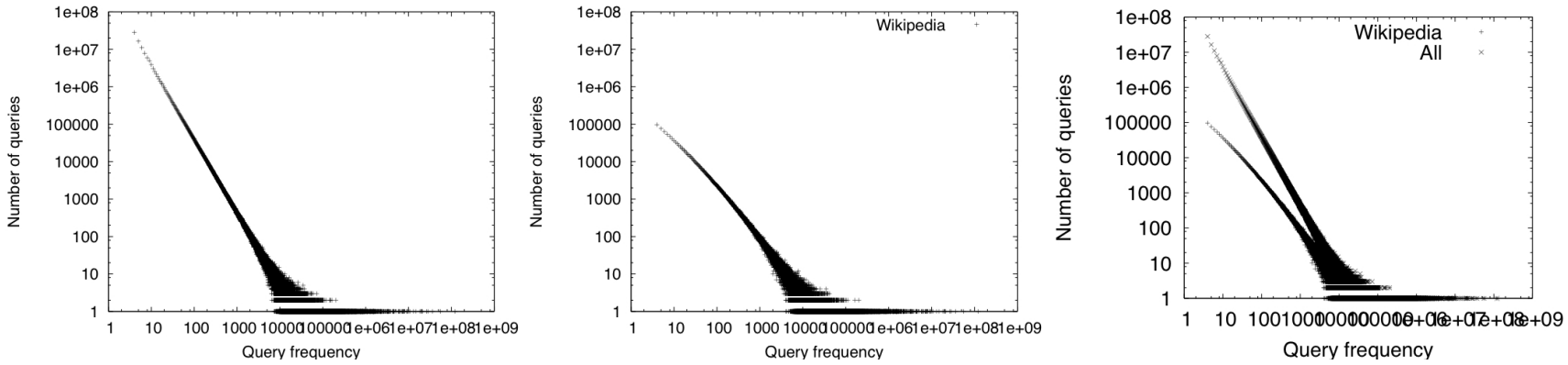
Wikipedia Coverage of Search Topics

- Do Wikipedia entries cover topics searched for by web users?
  - ⋆ We compare queries from a Web log with Wiki page titles
  - ⋆ 38.6% of 5.76 billion queries match Wikipedia page title

Wikipedia Coverage of Search Topics

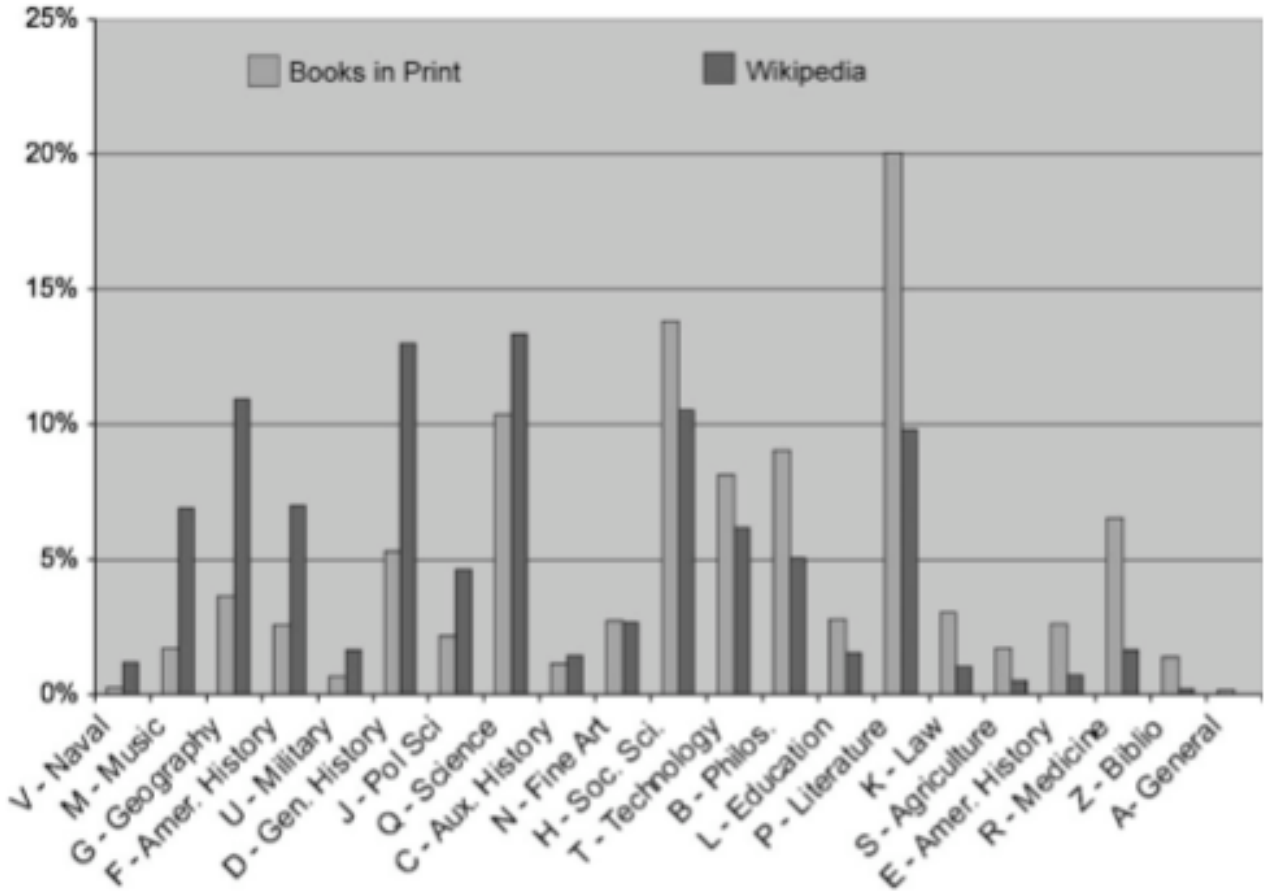
- Do Wikipedia entries cover topics searched for by web users?
  - ⋆ We compare queries from a Web log with Wiki page titles
  - ⋆ 38.6% of 5.76 billion queries match Wikipedia page title

Query frequency distribution



(a) All (b) Matching Wiki title (c) All & Matching Wiki title

Wikipedia coverage of book topics



- Source: Halavais and Lackaff (2008)
  - ⋆ Topics in published books and in sample of Wikipedia pages

Outline

- Introduction – Book Search
- Wikipedia Coverage
  - ⋆ Wikipedia coverage of search topics
  - ⋆ Wikipedia coverage of book topics
- **Wikipedia as intermediary**
  - ⋆ **Query Expansion**
  - ⋆ **Topical Closeness**
- Experiments & results
- Conclusion

Query Expansion

- Use Wiki page matching the query as rich topical description to draw terms from
  - ⋆ Using INEX Book Track corpus (42,095 books)
  - ⋆ Assumption: matching Wiki page is relevant

Query Expansion

- Use Wiki page matching the query as rich topical description to draw terms from
  - ⋆ Using INEX Book Track corpus (42,095 books)
  - ⋆ Assumption: matching Wiki page is relevant
- How to select terms?
  - ⋆ Terms from first paragraph
  - ⋆ Terms from anchor text
  - ⋆ Based on $tf.idf$ scores

Query Expansion

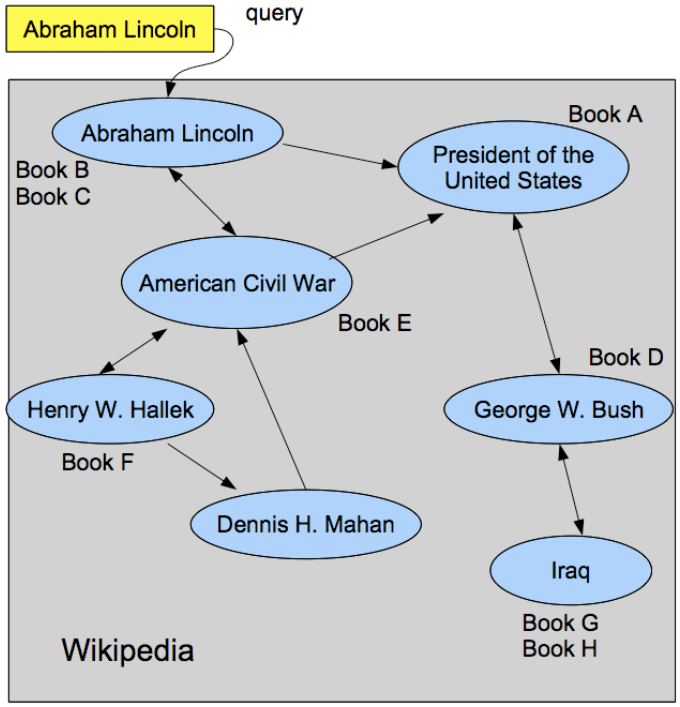
- Use Wiki page matching the query as rich topical description to draw terms from
  - ⋆ Using INEX Book Track corpus (42,095 books)
  - ⋆ Assumption: matching Wiki page is relevant
- How to select terms?
  - ⋆ Terms from first paragraph
  - ⋆ Terms from anchor text
  - ⋆ Based on $tf.idf$ scores
- Initial experiments show $tf.idf$ works best
  - ⋆ weight original query N times as much as the N added terms

Topical Closeness

- Wikipedia covers topics found in books (exit points):
  - ⋆ Users can traverse the link graph to related topics ($\rightarrow$ books)

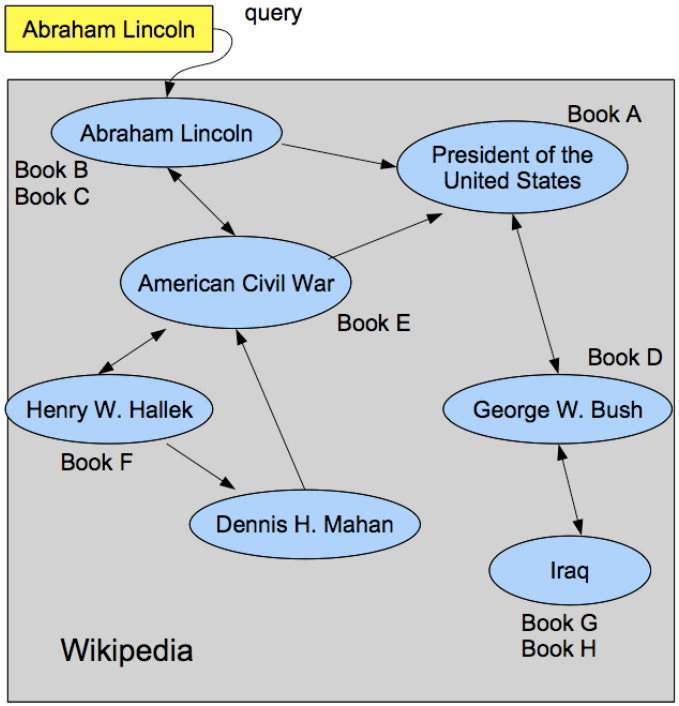
Topical Closeness

- Wikipedia covers topics found in books (exit points):
  - ⋆ Users can traverse the link graph to related topics (→ books)



Topical Closeness

- Wikipedia covers topics found in books (exit points):
  - ⋆ Users can traverse the link graph to related topics (→ books)



  - ⋆ Is link distance between search topics and book topics related to relevance?

Modelling Topical Closeness

- How can we associate books with topics in Wikipedia?

1. use book references on Wiki pages
2. use document similarity: book as query, rank Wiki pages

Modelling Topical Closeness

- How can we associate books with topics in Wikipedia?
  1. use book references on Wiki pages
  2. use document similarity: book as query, rank Wiki pages
- How can we measure the link distance between two topics in Wikipedia?
  - ⋆ Use random walk to compute closeness scores

Book References on Wikipedia Pages

References

- Basler, Roy P. (1946), *Abraham Lincoln: His Speeches and Writings.*
- Basler, Roy P. (1955), *Collected Works of Abraham Lincoln*, New Brunswick, NJ: Rutgers University Press
- Donald, David Herbert (1995), *Lincoln*, ISBN 0-684-82535-X.
- Foner, Eric (1970), *Free Soil, Free Labor, Free Men: The Ideology of the Republican Party before the Civil War*
- Jaffa, Harry V. (2000), *A New birth of Freedom: Abraham Lincoln and the Coming of the Civil War*, ISBN 0-8476-9952-8.
- Goodwin, Doris Kearns (2005), *Team of Rivals: The Political Genius of Abraham Lincoln*, ISBN 0-684-82490-6.
- Guelzo, Allen C. (1999), *Abraham Lincoln: Redeemer President*, ISBN 0-8028-3872-3
- Holzer, Harold (2004), *Lincoln at Cooper Union: The Speech That Made Abraham Lincoln President.*
- McPherson, James M. (1992), *Abraham Lincoln and the Second American Revolution.*
- Miller, William Lee (2002), *Lincoln's Virtues: An Ethical Biography*, ISBN 0-375-40158-X
- Sandburg, Carl (1974-10-23), *Abraham Lincoln: The Prairie Years and The War Years*, Harvest Books, ISBN 0156026112.
- Thomas, Benjamin P. (1952), *Abraham Lincoln: A Biography*.
- Wills, Garry (1993), *Lincoln at Gettysburg: The Words That Remade America*, ISBN 0-671-86742-3.
- Wilson, Douglas L. (1999), *Honor's Voice: The Transformation of Abraham Lincoln.*

- Many Wiki pages have references to books:
  - ⋆ Referenced books are relevant to the topic (?)

Book References on Wikipedia Pages

References

- Basler, Roy P. (1946), *Abraham Lincoln: His Speeches and Writings.*
- Basler, Roy P. (1955), *Collected Works of Abraham Lincoln*, New Brunswick, NJ: Rutgers University Press
- Donald, David Herbert (1995), *Lincoln*, ISBN 0-684-82535-X.
- Foner, Eric (1970), *Free Soil, Free Labor, Free Men: The Ideology of the Republican Party before the Civil War*
- Jaffa, Harry V. (2000), *A New birth of Freedom: Abraham Lincoln and the Coming of the Civil War*, ISBN 0-8476-9952-8.
- Goodwin, Doris Kearns (2005), *Team of Rivals: The Political Genius of Abraham Lincoln*, ISBN 0-684-82490-6.
- Guelzo, Allen C. (1999), *Abraham Lincoln: Redeemer President*, ISBN 0-8028-3872-3
- Holzer, Harold (2004), *Lincoln at Cooper Union: The Speech That Made Abraham Lincoln President.*
- McPherson, James M. (1992), *Abraham Lincoln and the Second American Revolution.*
- Miller, William Lee (2002), *Lincoln's Virtues: An Ethical Biography*, ISBN 0-375-40158-X
- Sandburg, Carl (1974-10-23), *Abraham Lincoln: The Prairie Years and The War Years*, Harvest Books, ISBN 0156026112.
- Thomas, Benjamin P. (1952), *Abraham Lincoln: A Biography*.
- Wills, Garry (1993), *Lincoln at Gettysburg: The Words That Remade America*, ISBN 0-671-86742-3.
- Wilson, Douglas L. (1999), *Honor's Voice: The Transformation of Abraham Lincoln.*

- Many Wiki pages have references to books:
  - ⋆ Referenced books are relevant to the topic (?)
- Small overlap with books in INEX collection (1,362 out of 42,095):
  - ⋆ Most books cited by Wiki pages published after 1970
  - ⋆ Most books in INEX corpus published up to 1930

Document Similarities

- Match each book in collection against Wiki page(s) based on document similarity

Document Similarities

- Match each book in collection against Wiki page(s) based on document similarity
- We indexed Wikipedia and used books as queries
  - ⋆ search engine can't handle whole book as query
  - ⋆ use the top 100 terms based on $tf.idf$ weights

Document Similarities

- Match each book in collection against Wiki page(s) based on document similarity
- We indexed Wikipedia and used books as queries
  - ⋆ search engine can't handle whole book as query
  - ⋆ use the top 100 terms based on $tf.idf$ weights
- Associate book with top N Wiki pages
  - ⋆ books can have multiple topics
  - ⋆ We experiment with $N = 1, 3, 5$
- All books in the INEX Book corpus can be matched

Computing Closeness

- We have linked both queries and books to Wiki pages.
- How to measure topical "closeness" in a graph?

Computing Closeness

- We have linked both queries and books to Wiki pages.
- How to measure topical "closeness" in a graph?
- Use random walk model:
  - ⋆ starting from page matching the query
  - ⋆ obtain closeness scores for all books

Computing Closeness

- We have linked both queries and books to Wiki pages.
- How to measure topical "closeness" in a graph?
- Use random walk model:
  - ⋆ starting from page matching the query
  - ⋆ obtain closeness scores for all books
- Are books found closer to the query topic more likely to be relevant than books further away from it?

Closeness and Probability of Relevance

- We can compute the probability of relevance (PoR) over topical closeness

Closeness and Probability of Relevance

- We can compute the probability of relevance (PoR) over topical closeness
- Each score represents closeness between a query and a book
  - ⋆ sort scores and bin per 10,000 scores

Closeness and Probability of Relevance

- We can compute the probability of relevance (PoR) over topical closeness
- Each score represents closeness between a query and a book
  - ⋆ sort scores and bin per 10,000 scores
  - ⋆ count scores representing a query and a book relevant to that query

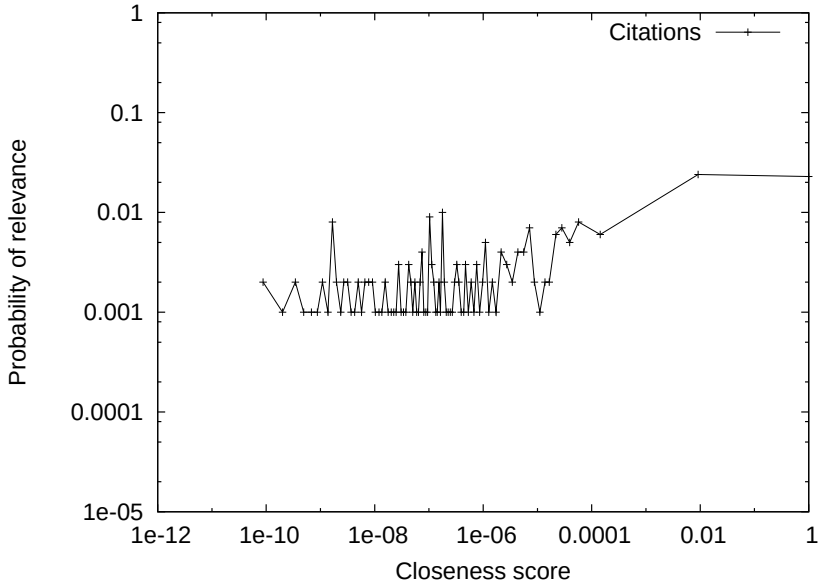
Closeness and Probability of Relevance

- We can compute the probability of relevance (PoR) over topical closeness
- Each score represents closeness between a query and a book
  - ⋆ sort scores and bin per 10,000 scores
  - ⋆ count scores representing a query and a book relevant to that query
  - ⋆ PoR is the ratio of relevant scores in each bin

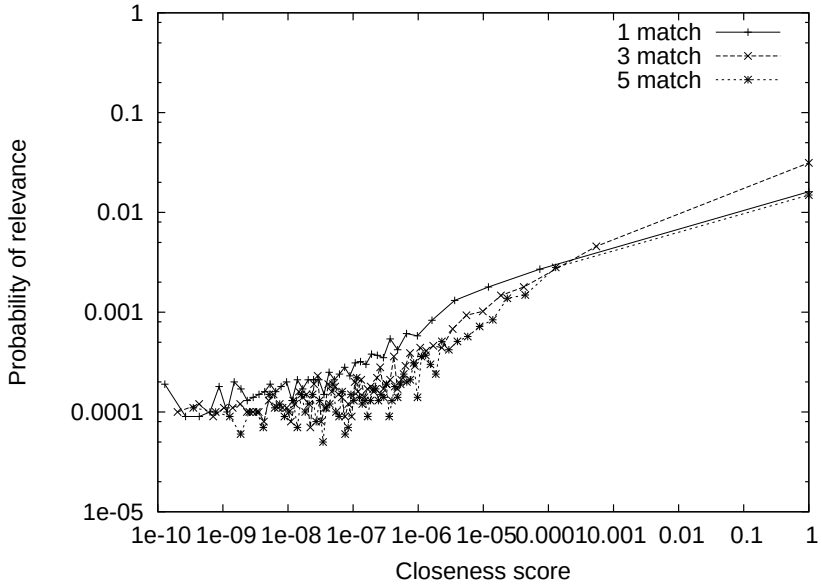
Closeness and Probability of Relevance

- We can compute the probability of relevance (PoR) over topical closeness
- Each score represents closeness between a query and a book
  - ⋆ sort scores and bin per 10,000 scores
  - ⋆ count scores representing a query and a book relevant to that query
  - ⋆ PoR is the ratio of relevant scores in each bin
- If closeness is related to relevance, we expect PoR to go up with increasing closeness score

Closeness and Relevance

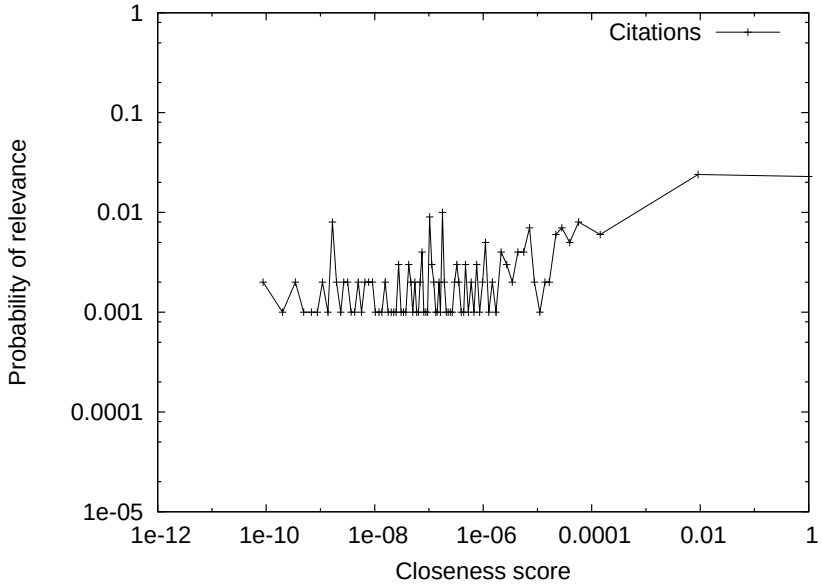


References

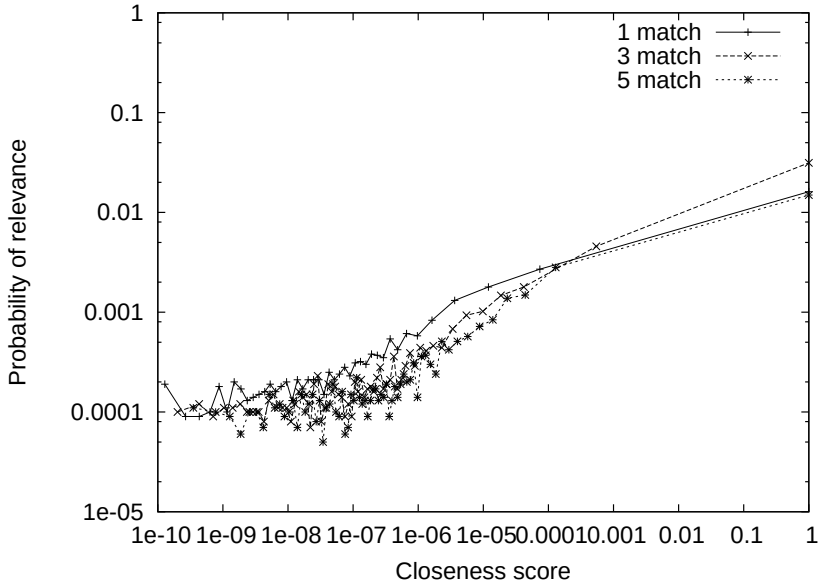


Doc. sim.

Closeness and Relevance



References



Doc. sim.

- We see:
  - ⋆ Only at higher scores (> 0.0001) do we see a rising trend
  - ⋆ Document similarity seems the more stable indicator

Outline

- Introduction
- Wikipedia Coverage
  - ⋆ Wikipedia coverage of search topics
  - ⋆ Wikipedia coverage of book topics
- Wikipedia as Intermediary
  - ⋆ Query expansion
  - ⋆ Topical Closeness
- **Experiments & Results**
- Conclusion

Experiments

- Books indexed using Lemur/Indri
- INEX 2007 Book corpus
  - ⋆ 42,095 books
  - ⋆ 250 topics with relevance judgements from Live Search
  - ⋆ On average, 15.56 judgements per query
- 176 queries match title of a Wiki page (70.4%)

Query Expansion

Run id	# jugded rel.	# jugded non-rel.	MAP	Bpref	P10
baseline	1666	808	0.3771	0.6131	0.3040
$N = 5$	1666	808	0.3725	**0.6205**	0.3080
$N = 10$	1671	808	**0.3874***	0.6168	0.3119*
$N = 20$	1667	807	0.3837*	0.6149	**0.3136***
$N = 50$	1666	806	0.3780	0.6136	0.3074*
$N = 100$	1666	807	0.3780*	0.6133	0.3063*

Query Expansion

Run id	# jugded rel.	# jugded non-rel.	MAP	Bpref	P10
baseline	1666	808	0.3771	0.6131	0.3040
$N = 5$	1666	808	0.3725	**0.6205**	0.3080
$N = 10$	1671	808	**0.3874***	0.6168	0.3119*
$N = 20$	1667	807	0.3837*	0.6149	**0.3136***
$N = 50$	1666	806	0.3780	0.6136	0.3074*
$N = 100$	1666	807	0.3780*	0.6133	0.3063*

- We see:
  - $\star$ QE has almost no effect on number of relevant documents

Query Expansion

Run id	# jugded rel.	# jugded non-rel.	MAP	Bpref	P10
baseline	1666	808	0.3771	0.6131	0.3040
$N = 5$	1666	808	0.3725	**0.6205**	0.3080
$N = 10$	1671	808	**0.3874***	0.6168	0.3119*
$N = 20$	1667	807	0.3837*	0.6149	**0.3136***
$N = 50$	1666	806	0.3780	0.6136	0.3074*
$N = 100$	1666	807	0.3780*	0.6133	0.3063*

- We see:
  - ⋆ QE has almost no effect on number of relevant documents
  - ⋆ improvements are small but significant for MAP and P10
  - ⋆ impact drops with increasing N (consequence of term weighting)

Topical Closeness

Run id	MAP	Bpref	P10
baseline	0.3771	0.6131	0.3040
References	0.3769	0.6150	0.3051
Doc.Sim.1	0.3604	0.6010	0.2983
Doc.Sim.3	0.3790	0.6245*	**0.3091***
Doc.Sim.5	**0.3823***	**0.6251***	0.3080*

- Final RSV is Indri score + sigmoid trans. of closeness score

Topical Closeness

Run id	MAP	Bpref	P10
baseline	0.3771	0.6131	0.3040
References	0.3769	0.6150	0.3051
Doc.Sim.1	0.3604	0.6010	0.2983
Doc.Sim.3	0.3790	0.6245*	**0.3091***
Doc.Sim.5	**0.3823***	**0.6251***	0.3080*

- Final RSV is Indri score + sigmoid trans. of closeness score
  - ⋆ References have small impact (because of small overlap)
  - ⋆ Doc.Sim. using top 1 Wiki page hurts performance

Topical Closeness

Run id	MAP	Bpref	P10
baseline	0.3771	0.6131	0.3040
References	0.3769	0.6150	0.3051
Doc.Sim.1	0.3604	0.6010	0.2983
Doc.Sim.3	0.3790	0.6245*	**0.3091***
Doc.Sim.5	**0.3823***	**0.6251***	0.3080*

- Final RSV is Indri score + sigmoid trans. of closeness score
  - ⋆ References have small impact (because of small overlap)
  - ⋆ Doc.Sim. using top 1 Wiki page hurts performance
  - ⋆ Doc.Sim. using multiple Wiki pages improves all measures (significantly for $N = 5$)

Conclusions (1/2)

- Wikipedia as intermediary between user and book collections:
  - ⋆ Wikipedia covers many search topics and books topics

Conclusions (1/2)

- Wikipedia as intermediary between user and book collections:
  - ⋆ Wikipedia covers many search topics and books topics
- Can we automatically extract useful terms from related Wikipedia pages to improve retrieval effectiveness?
  - ⋆ Yes, QE using $tf.idf$ term selection from single entry point leads to small improvements
  - ⋆ Problem: entry point might not be relevant

Conclusions (2/2)

- Is the link distance between query pages and book pages related to relevance and can we use this to improve retrieval effectiveness?
  - ⋆ link distance shows weak relation to relevance
  - ⋆ document similarity using multiple Wiki pages can significantly improve performance

Conclusions (2/2)

- Is the link distance between query pages and book pages related to relevance and can we use this to improve retrieval effectiveness?
  - ⋆ link distance shows weak relation to relevance
  - ⋆ document similarity using multiple Wiki pages can significantly improve performance
- Low number of judgements might not properly reflect effectiveness of chosen methods
  - ⋆ Last year's (2008) INEX Book Track topics have deeper pools
  - ⋆ judgements are about to be released

Future Work

- Use different book representations to find best matching Wikipedia pages:
  - ⋆ Use collocations, latent semantic indexing

Future Work

- Use different book representations to find best matching Wikipedia pages:
  - ⋆ Use collocations, latent semantic indexing
- There are uninformative links (Abraham Lincoln $\longrightarrow$ 2006)
  - ⋆ leads to noise closeness scores
  - ⋆ filter uninformative links
  - ⋆ weight links by measuring document similarity between two topics

Thank You!