IR in Practice:
Patent Retrieval

Parvaz Mahdabi
Advisor: Prof. Fabio Crestani

IR group, University of Lugano, Switzerland

Università della Svizzera italiana

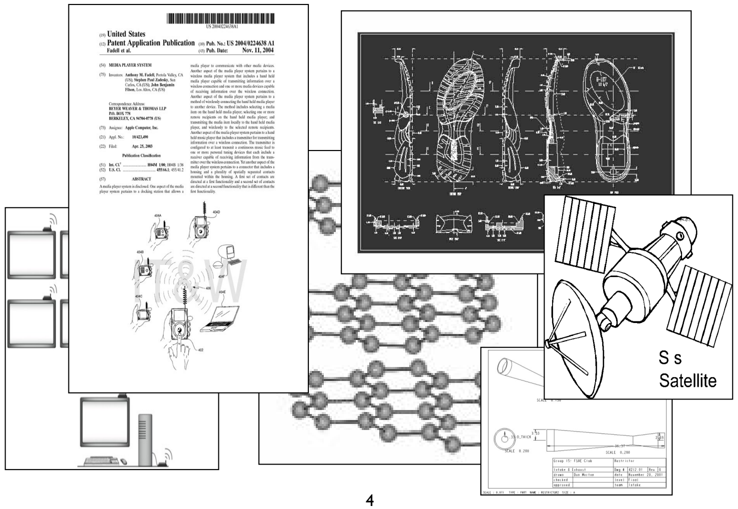
Faculty of Informatics

Nov 2013

Outline

1. What is Patent Retrieval?
2. Differences with Standard IR
3. Challenges
4. Related Work
5. Building Query from Patent
6. Domain-dependent Lexicon
7. Query Expansion using Proximity clues
8. Conclusions

What is Patent and Patent Searching?



What is a Patent?

- An official document, issued by a Patent office, granting property rights to the inventor or assignee and the right to EXCLUDE others from making, using, offering for sale, selling or importing the invention.
- Term is generally 20 years from the date of application in the U.S, if maintenance fees are paid.
- The first to file a patent is the inventor (gets the credit)

What Inventions can be Patented?

- A **new** and **useful**
  - process
  - machine
  - article of manufacture
  - composition of matter
  - or any useful and new improvements on the above

Requirement of "USEFUL"

- The invention has a useful purpose
- The invention will perform to operate the useful purpose, *i.e. it works.*

Requirement of "NEW"

- The invention has not been disclosed before (novelty)
  - public disclosure includes written (article), verbal (conference presentation), sale, or offer for sale (marketing)

Why Search for Patents?

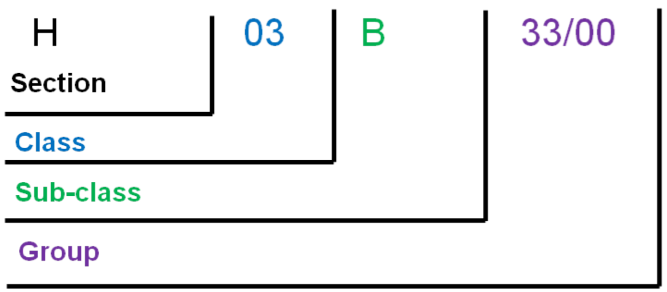
- New and innovative technologies
- Competitive intelligence
- Background on technologies not covered in journal/conference articles
- Patentability
- ....

The Patent Application

- Title
- Description of invention
- One or more claims which are carefully worded statements to determine the boundaries of the invention
- Drawings if necessary

IPC classes
(Hierarchical Classification System)

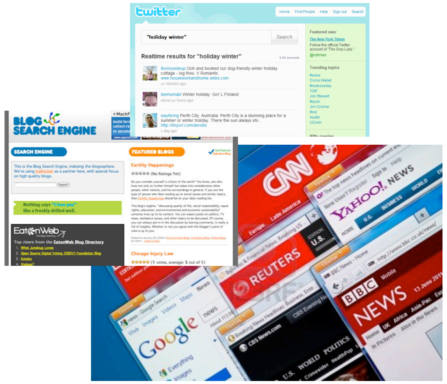
H03B33/00



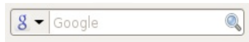
Hierarchical Structure of IPC classes

H	ELECTRICITY
H03	BASIC ELECTRONIC CIRCUITRY
H03B	GENERATION OF OSCILLATIONS, DIRECTLY OR BY FREQUENCY-CHANGING, BY CIRCUITS EMPLOYING ACTIVE ELEMENTS WHICH OPERATE IN A NON-SWITCHING MANNER; GENERATION OF NOISE BY SUCH CIRCUITS ...
H03B 5/04	· · Modifications of generator to compensate for variations in physical values, e.g. power supply, load, temperature

Patent Retrieval Versus Standard Information Retrieval



Web Search



User
- normal user

Query
- Short (2, 3 words)

Goal of search
- Precision oriented
- The first page is examined

Prior-art Search

Title <ttl> ... </ttl>

Abstract<abs> ... </abs>

Description<dsc> ... </dsc>

Claims<clm> ... </clm>

IPC <ipc> ... </ipc>

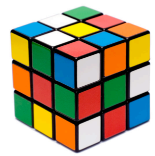
User
- Expert user (Patent Examiner)

Query
- Full application (hundreds of words)

Goal of search
- Recall oriented
- A long list is examined

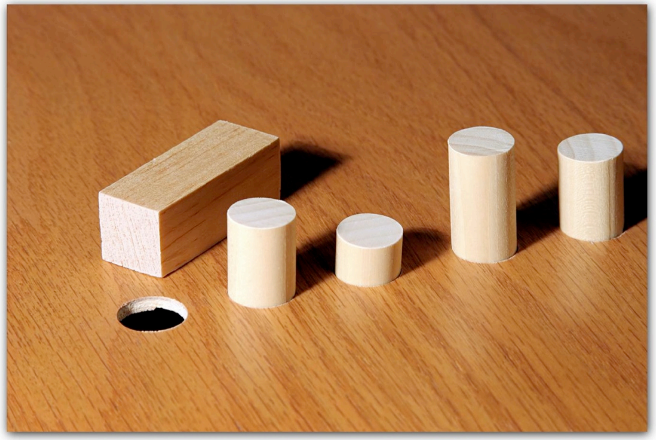
Check Novelty

Challenges of Prior-art Search

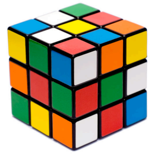


- A full patent application instead of a keyword query
  - Incorporating different relevance evidences such as textual content, patent classification, bibliographic information, publication dates, ...
- Legal terminology (different set of stop-words)
- Recall-oriented (satisfy legal requirements)

Query Document Mismatch is biggest challenge in Patent Retrieval



Challenges of Prior-art Search



- Significant term mismatch (Query: "ipod", Document= "music player")
  - Usage of new inventive words
  - Rewording (for avoiding repetition)
  - Non-standardized acronyms: invented by authors
  - Synonyms: signal and wave
  - Homonyms: bus (1- motor vehicle, 2- within a computer system)

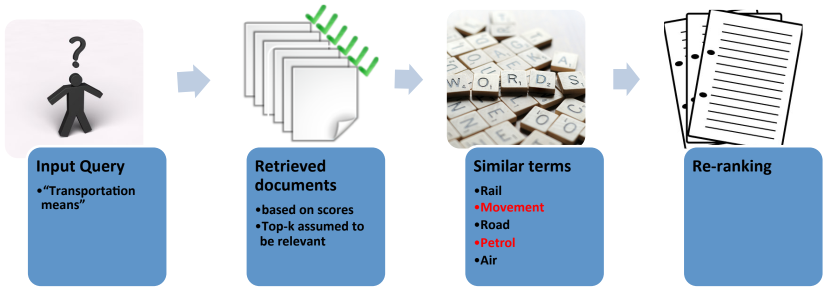
Query Document Matching at Different Levels

Level of Semantics

Structure	how green is the technology → the fuel consumption rate
Topic	transportation vehicle → ..., bicycle, car, bus, ...
Word Sense	the interior of an object → inside
Phrase	"transportation vehicle" → "transportation vehicle"
Term	transportation → transportation

Standard Pseudo Relevance Feedback

for Minimizing Term Mismatch



Sta ard Pseudo Rele back

Does not perform well on Patent data

Standard PRF is ineffective for patent retrieval due to the low precision of the original rank list [Ganguly et al, 2011]

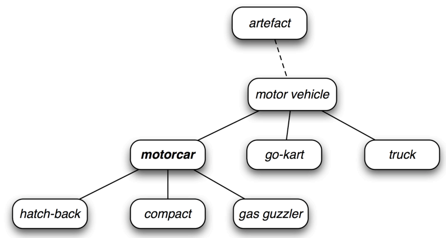
means"

•based on scores
•Top-k assumed to be relevant

•Movement
•Road
•Petrol
•Air

Query Expansion

for Minimizing Term Mismatch



WordNet

Synonyms

WIKIPEDIA
The Free Encyclopedia

Topically Relevant terms
Disambiguation pages

Query Resources
for Min

perform marginally successful on Patent data, still not comparable to news text

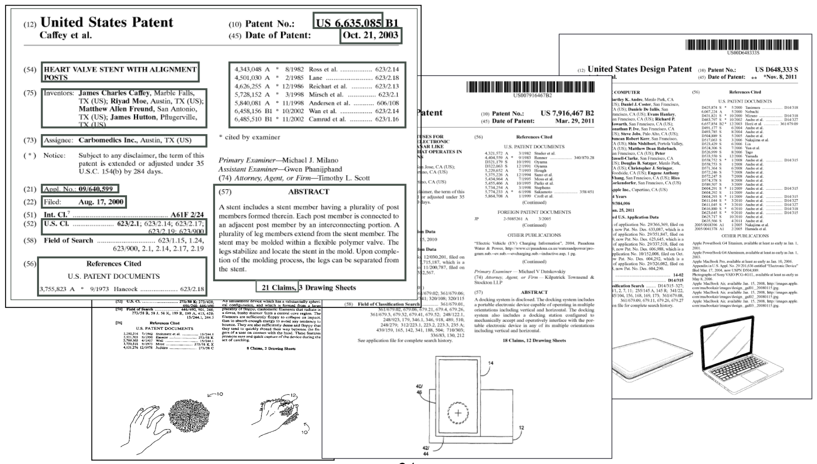
artefact

Use of synonyms in WordNet for Patent Retrieval is not effective for improving recall (Magdy and Jones, 2011)

Successful Exploitation of Wikipedia information for query expansion (Lopez et al., 2010)

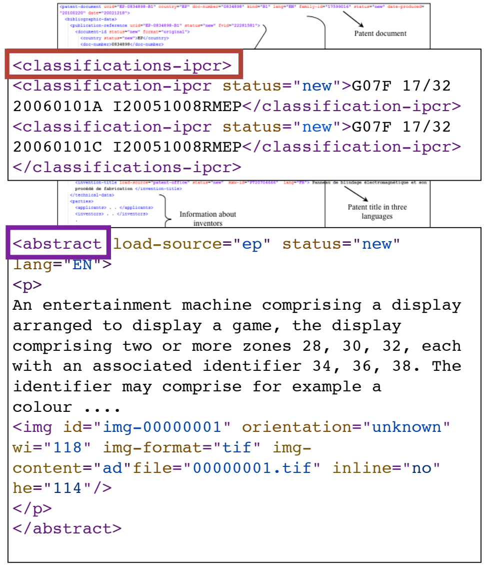
and use for query expansion

Patent Collection used in our Experiments



Patent Document

- Patent classifications
- Inventor information
- Title
- Abstract
- Description
- Claims



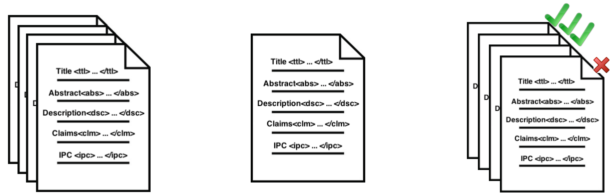

```xml
<classifications-ipcr>
<classification-ipcr status="new">G07F 17/32
20060101A I20051008RMEP</classification-ipcr>
<classification-ipcr status="new">G07F 17/32
20060101C I20051008RMEP</classification-ipcr>
</classifications-ipcr>

<abstract load-source="ep" status="new"
lang="EN">
<p>
An entertainment machine comprising a display
arranged to display a game, the display
comprising two or more zones 28, 30, 32, each
with an associated identifier 34, 36, 38. The
identifier may comprise for example a
colour ....
<img id="img-00000001" orientation="unknown"
wi="118" img-format="tif" img-
content="ad"file="00000001.tif" inline="no"
he="114"/>
</p>
</abstract>
```

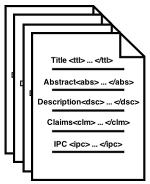
Downloadable at: http://ifs.tuwien.ac.at/~clef-ip/

Test Collection for Experiments

Collection | Query Patent | Relevance Judgements



Patent Collection



- CLEF-IP 2010
  - 1.3 million patent documents (unzipped: 100 Gig)
  - contains granted patent applications from 1976 to 2008
- training set (for tuning parameters)
- test set (used for performance comparison among methods)

Our Proposed Solution

1. Query Generation from Patent Application

Building queries for prior art search [Mahdabi et al, 2011]

2. Query Expansion using conceptual lexicon

Leveraging Conceptual Lexicon: Query Disambiguation using Proximity Information for Patent Retrieval [Mahdabi et al, 2013]

Related Work on Query Generation

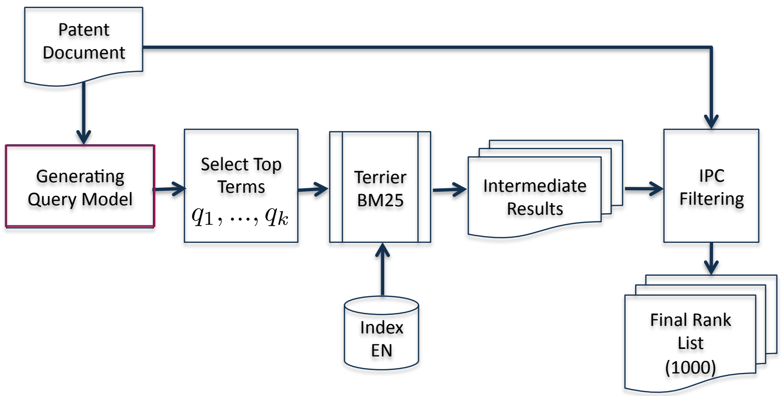
- Reducing patent query
  - using learning to rank approaches [Xue et al., 2010]
  - using conditional random field and exploit Wikipedia information [Lopez et al., 2010]
  - using proximity information [Bashir et al., 2010]
  - query reduction using pseudo relevant documents [Ganguly et al., 2011]
- Evaluation metric for patent retrieval
  - PRES: combines Recall and Precision, importance to recall [Magdy et al., 2010]

Query Generation

- Identify important terms
- Which sections should be used
- Estimate the query model in a Language Modeling framework

Building queries for prior art search [Mahdabi et al, 2011]

System Architecture



Terrier: http://terrier.org/

Step 1 compute term distribution

1. Maximum likelihood estimate

$$P_{ML}(w|Q_f) = \frac{tf(w, Q_f)}{|Q_f|}$$

Query Document

$$P_{ML}(w|Cluster_f) = \frac{1}{N} \sum_{D_f \in RIPC} \frac{tf(w, D_f)}{|D_f|}$$

Cluster of documents with common IPC classes as Query Document

Step 2
smoothing

2. Cluster smoothed estimate (Liu et al., 2004)

$$P(t|\theta_{Q_f}) = \lambda P_{ML}(t|Q_f) + (1-\lambda) P_{ML}(t|Cluster_f)$$

Smoothing parameter Document model Cluster model

Step 3
rank terms

a. Rank terms based on LLQM

their high similarity to the document
and low similarity to the corpus

$D_{KL}(p(w|\theta_{Q_f}))||p(w|\theta_{Coll_f})$

b. Rank terms based on CBQM

their high similarity to the document and the cluster
and low similarity to the corpus

$H(\theta_{Q_f}, \theta_{Coll_f}) - H(\theta_{Q_f}, \theta_{Cluster_f})$

Kullback–Leibler divergence

- is a non-symmetric measure of the difference between two probability distributions *P* and *Q*
- the Kullback–Leibler divergence of *Q* from *P*, denoted $D_{KL}(P \| Q)$, is a measure of the information lost when *Q* is used to approximate *P*

$$D_{KL}(P||Q) = -\sum_x p(x) \log q(x) + \sum_x p(x) \log p(x)$$

$$= \underbrace{H(P,Q)}_{\text{Cross Entropy}} - \underbrace{H(P)}_{\text{Entropy}}$$

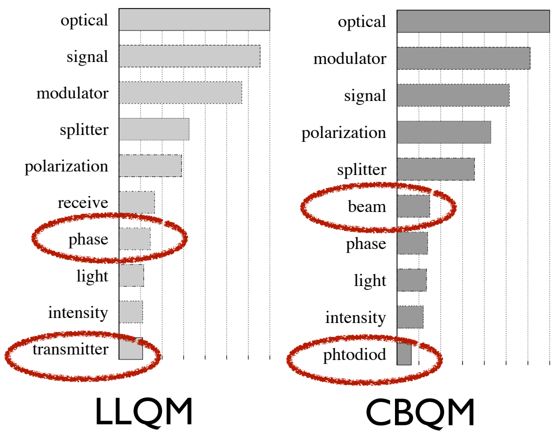
The effect of the term source

Training data is used to set the parameters
Performance results are reported on the test set

LLQM	MAP	Recall	PRES
Title	0.05	0.53	0.42
Abstract	0.07	0.56	0.45
Claim	0.10	0.57	0.47
Description	0.12	0.63	0.50
All Text	0.09	0.57	0.47

top-10 query terms extracted from patent application

"System and method for multi-level phase modulated communication"



Related Work on Query Expansion

Address term mismatch using external resources

- Use of Wikipedia by Lopez and Romary (CLEF 2010)
- Use of WordNet by Magdy and Jones (CIKM 2011)

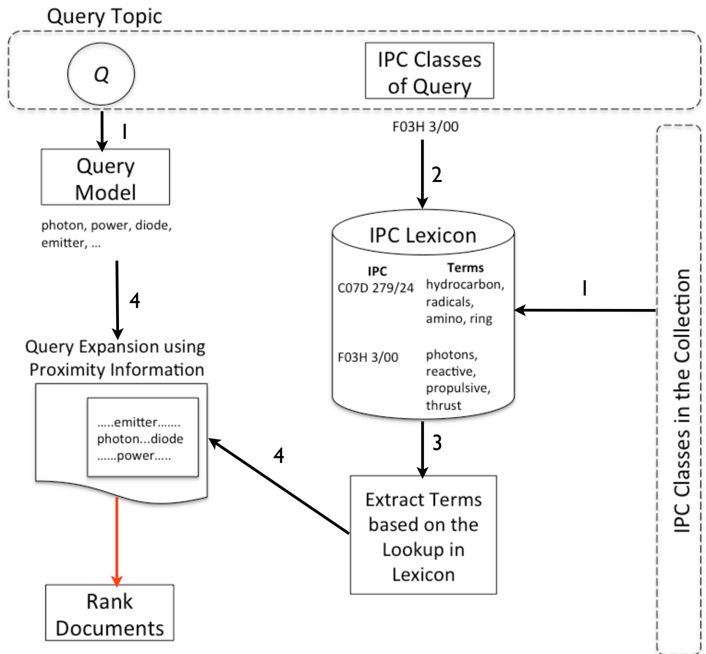
Using proximity evidences

- Use of passages to capture term positions by Ganguly et al (CIKM 2011)
- Use proximity heuristics (distance of query term to expansion term) for query expansion by Bashir and Rauber (ECIR 2010)

Related Work

Use of proximity information in a systematic way in IR

- "positional language model" and "positional relevance model" by Lv and Zhai (SIGIR 2009, SIGIR 2010)
- Capturing opinion density for improving blog retrieval by Gerani et al (SIGIR 2010)



Building Domain-dependent Lexicon

- Our conceptual lexicon is based on explanation of IPC classes

IPC Class	Definition
C07D 279/24	· · · · · with hydrocarbon radicals, substituted by amino radicals, attached to the ring nitrogen atom

Building Domain-dependent Lexicon

- Stop word removal on the text of IPC definition pages
- Increase the accuracy by filtering out patent-specific stop-words ("method", "device", "apparatus", "process")
- Each entry in the lexicon is composed of a key and a value

IPC Class	Representing Terms
C07D 279/24	hydrocarbon, radicals, amino, ring, nitrogen, atom

Assumptions

1. An expansion term refer with higher probability to the query terms closer to its position (proximity operators are used in the real task of patent examiners, NEAR, ADJ)

   - We model the query term influence propagation with density kernel functions

Proximity is used

Assumptions

2. A query term might belong to

   - the author terminology
   - the vocabulary of IPC classes
   - the vocabulary of the community of inventors (cited documents)

Author and IPC classes are used in query formulation

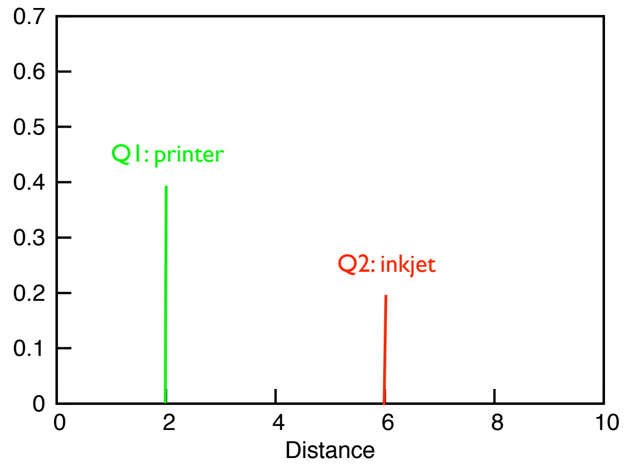
Kernel Density Functions

- A non-parametric way to estimate the probability density function of a random variable
- The probability density function is nonnegative everywhere, and its integral over the entire space is equal to one
- The probability of a random value falling in a range is given by the area under the density function between the lowest and greatest values of the range

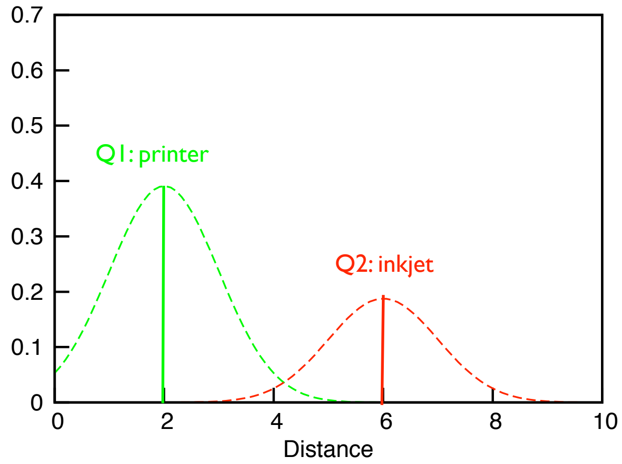
Modeling Term Dependency with Kernel Functions

- Lifting probability mass around query term occurrence, so that adjacent terms receive higher probability

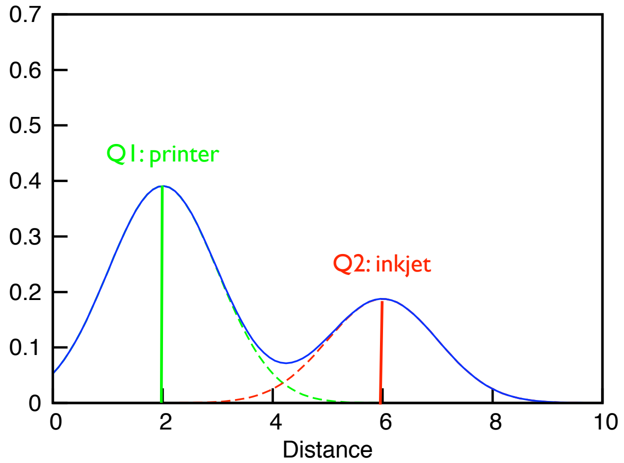
Query Relatedness Density *P(q|i,d)*



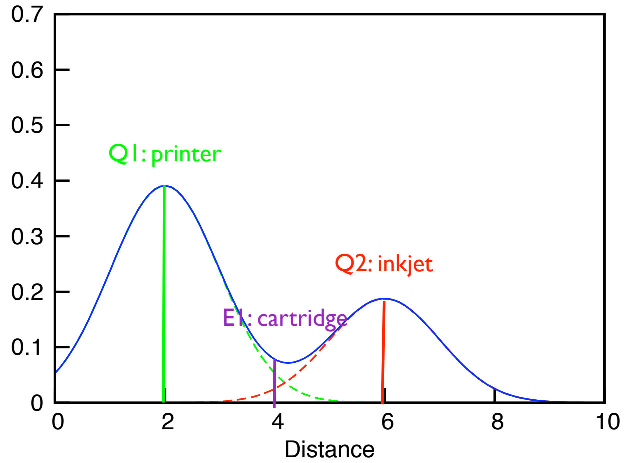
Query Relatedness Density *P(q|i,d)*



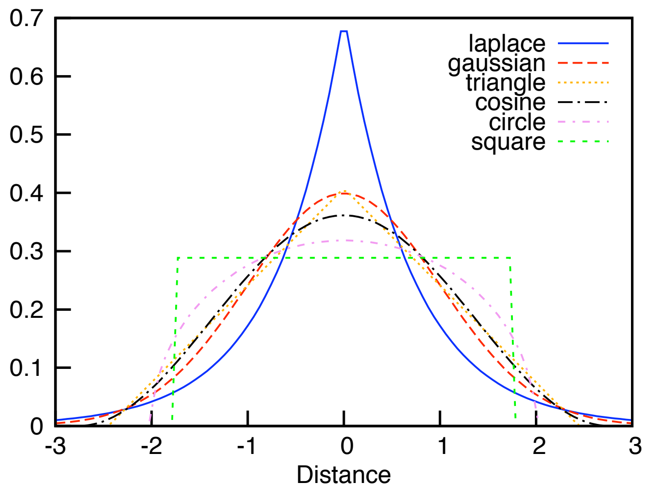
Propagated Query Relatedness



Propagated Query Relatedness



Kernel Density Functions



Building the Initial Query

$$P(t|\theta_{Orig}) = Z_t P(t|\theta_Q) log(\frac{P(t|\theta_Q)}{P(t|\theta_C)})$$

$P(t|\theta_Q)$ query language model

$P(t|\theta_C)$ collection language model

Z_t normalization factor

Calculating Document Relevance Score

- Overall probability that relevant expansion terms (inside the document) are directed towards the technical concept of the query

$$P(q|d,e) = \sum_{i=1}^{|d|} \underbrace{P(q|i,d,e)}_{\text{Query-relatedness}} \underbrace{P(i|d,e)}_{\text{Expansion}}$$

Estimating the Query Relatedness

- Assume e and *q* are conditionally independent given the position in the *d* thus *P(q|i,d,e)* reduces to *P(q|i,d)*
- Estimate the probability that an expansion term e at position *i*, is related to the query term *q* at position *j*

$$\underbrace{P(q|i,d)}_{\text{query-relatedness}} = \sum_{j=1}^{m} \underbrace{P(q|t_j)}_{\text{query weight}} \underbrace{P(j|i,d)}_{\text{Proximity-based estimate}}$$

- $P(q|t_j)$: comes from the initial query model

Estimating the Query Relatedness

$$\underbrace{P(j|i,d)}_{\text{Proximity-based estimate}} = \frac{k(j,i)}{\sum_{j'=1}^{|d|} k(j',i)}$$

- $P(j|i,d)$ is formed by placing a density kernel function around each query term
- $k(j,i)$ is a kernel function which determines the weight of propagated query relatedness from t_j to t_i

Estimating the Expansion Probability

- Avg Strategy: All positions of expansion terms are equally important

$$\begin{cases} P(i|d,e) = 1/|pos(e)| & \text{if } t_i \in e \\ 0 & \text{otherwise} \end{cases}$$

$$P(q|d,e) = 1/|pos(e)| \sum_{i \in pos(e)} P(q|i,d)$$

- Max Strategy: The expansion position with the maximum probability is important

$$P(q|d,e) = \max_{i \in pos(e)} P(q|i,d)$$

Experimental Settings

- Language Modeling with Dirichlet smoothing is used to score documents in the initial rank lists
- Terrier* is used for building the index
- CLEF-IP 2010 training set is used for tuning the parameters

*Terrier: http://terrier.org/

Recall Results of Different Settings of Kernel Functions

Query Expansion				
Kernel\ sigma	25	75	125	150
Gaussian	**0.6443**	**0.6561**	**0.6676**	**0.6795**
Laplace	0.6422	0.6556	0.6588	0.6709
Square	0.6398	0.6523	0.6559	0.6678

Simulate Passage Retrieval

Conclusions

- Patent specific stop words are different from standard text (news)
- Proximity information is important in patent retrieval
- A domain dependent lexicon built form patent classifications is more effective for query expansion compared to using Wikipedia or WordNet
- Kernel density function are used to model dependency between words

IR in Practice:
Patent Retrieval

Parvaz Mahdabi
parvaz.mahdabi@usi.ch

Nov 2013

References

S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In ECIR, pages 457-470, 2010.

D. Ganguly, J. Leveling, W. Magdy, and G. J. F. Jones. Patent query reduction based on pseudo-relevant documents. In CIKM, pages 1953-1956, 2011.

S. Gerani, M. J. Carman, and F. Crestani. Aggregation methods for proximity-based opinion retrieval. TOIS, 30(4):26, 2012.

P. Lopez and L. Romary. Patatras: Retrieval model combination and regression models for prior art search. In CLEF (Notebook Papers/LABs/Workshops), pages 430-437, 2009.

P. Lopez and L. Romary. Experiments with citation mining and key-term extraction for prior art search. CLEF (Notebook Papers/LABs/Workshops), 2010.

Y. Lv and C. Zhai. Positional language models for information retrieval. In SIGIR, pages 299-306, 2009.

References

Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In SIGIR, pages 579-586, 2010.

W. Magdy and G. J. F. Jones. PRES: A score metric for evaluating recall-oriented information retrieval applications. In SIGIR, pages 611-618, 2010.

W. Magdy and G. J. F. Jones. A study on query expansion methods for patent retrieval. In PAIR 2011, CIKM, pages 19-24, 2011.

P. Mahdabi, M. Keikha, S. Gerani, M. Landoni, F. Crestani: Building Queries for Prior-Art Search. In IRFC 2011, pages 3-15

P. Mahdabi, S. Gerani, J. Huang, F. Crestani, "Leveraging Conceptual Lexicon: Query Disambiguation using Proximity Information for Patent Retrieval" , In SIGIR, pages 113-122, 2013.

X. Xue and W. B. Croft. Automatic query generation for patent search. CKIM, pages 2037-2040, 2009.