Monsoon 2020

12 - Query Expansion

Information Retrieval

by

Dr. Rajendra Prasath

Indian Institute of Information Technology, Sri City, Chittoor

Sri City – 517 646, Andhra Pradesh, India

✧ Topics Covered So Far

- ✧ Bi-Word Index / Wildcard Queries / Permuterm Index
- ✧ K-gram Index (k = 2 □ Bigram Index)
- ✧ Spell Correction
- ✧ Term Weighting
- ✧ Vector Space Models
- ✧ IR Evaluation Metrics
- ✧ Relevance Feedback Approaches
- ✧ Pseudo Relevance Feedback Approaches

✧ Now:

Query Expansion Approaches

Recap:

- Why Ranked Retrieval?
- Term Frequency
- Term Weighting
- TF-IDF Weighting
- The Vector Space Model
- Relevance Feedback
- Pseudo Relevance Feedback

Recap: Vector Space Models

- The length of the sub-vector in dimension - i is used to represent the importance or the weight of the word – i in a text
- Words that are absent in a text get a weight zero

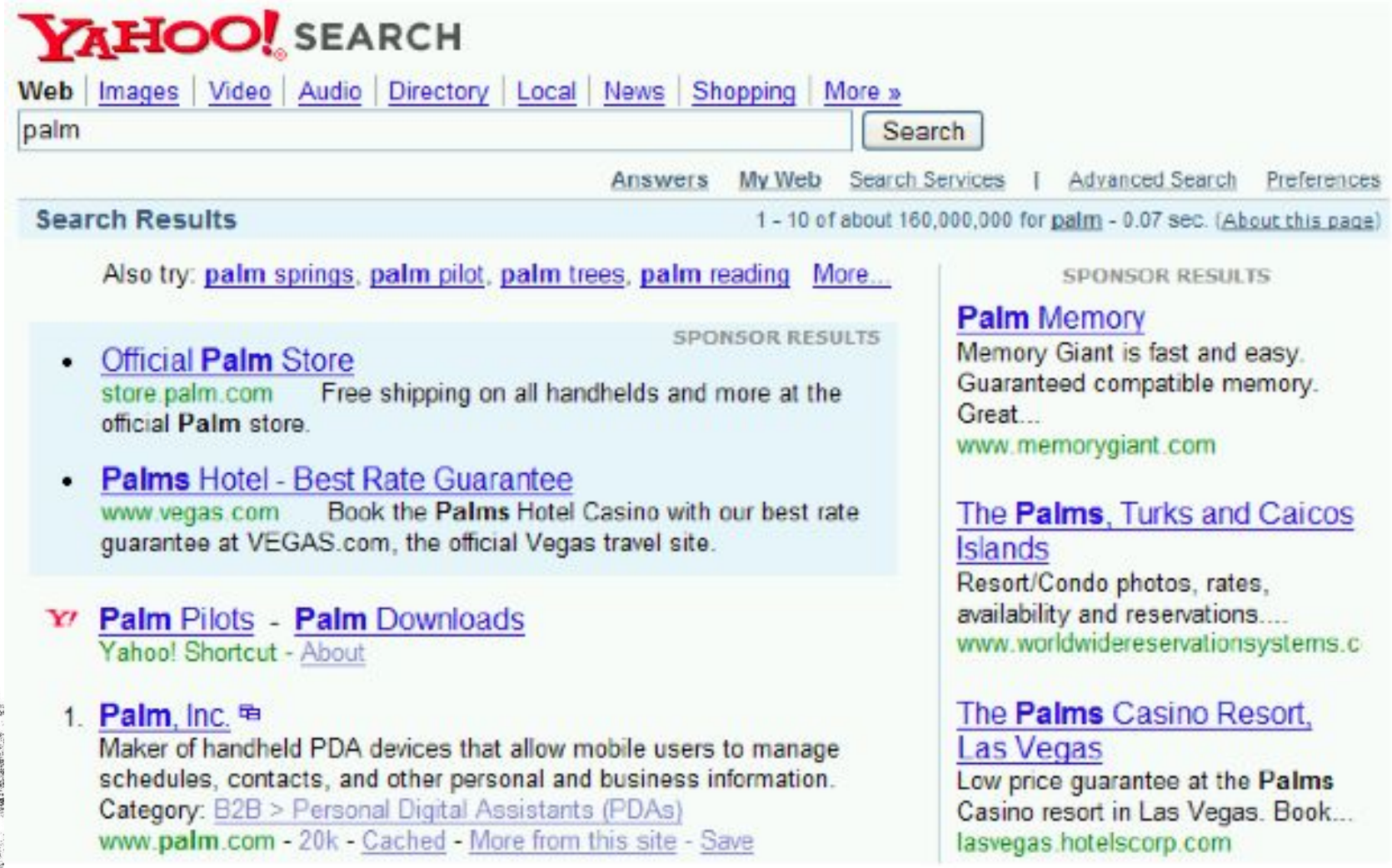
- Apply **Vector Inner Product** measure between two vectors:
- This vector inner product increases:
  - # words match between two texts
  - Importance of the matching terms

Query Expansion

Query expansion

- Another Way to increase recall
- Global query expansion
  ⇒ global methods for query reformulation
- In global query expansion, the query is modified based on some global resource, i.e. a resource that is not query-dependent.
- Main information we use: (near-)synonymy
- A publication or database that collects (near-)synonyms is called a thesaurus.
- We will look at two types of thesauri: manually created and automatically created

Query expansion: Example



Types of user feedback

- User gives feedback on documents.
  - More common in relevance feedback
- User gives feedback on words or phrases.
  - More common in query expansion

Types of query expansion

- Manual thesaurus (maintained by editors, e.g., PubMed)
- Automatically derived thesaurus (e.g., based on co-occurrence statistics)
- Query-equivalence based on query log mining (common on the web as in the “palm” example)

Thesaurus-based query expansion

- For each term t in the query, expand the query with words the thesaurus lists as semantically related
- Example: HOSPITAL → MEDICAL
- Generally increases recall
- May significantly decrease precision, particularly with ambiguous terms
- INTEREST RATE → INTEREST RATE FASCINATE
- Widely used in specialized search engines for science and engineering
- Expensive to create/maintain a manual thesaurus
- A manual thesaurus has an effect roughly equivalent to annotation with a controlled vocabulary.

Automatic thesaurus generation

- Attempt to generate a thesaurus automatically by analyzing the distribution of words in documents
- Fundamental notion: similarity between two words

Two words are similar if they co-occur with similar words

- "car" ≈ "motorcycle" because both occur with "road", "gas" and "license", so they must be similar.

Two words are similar if they occur in a given grammatical relation with the same words

- You can harvest, peel, eat, prepare, etc. apples and pears, so apples and pears must be similar.
- Co-occurrence - more robust; Grammatical relations - more accurate.

Co-occurence-based thesaurus: Examples

Word	Nearest neighbors
absolutely	absurd whatsoever totally exactly nothing
bottomed	dip copper drops topped slide trimmed
captivating	shimmer stunningly superbly plucky witty
doghouse	dog porch crawling beside downstairs
makeup	repellent lotion glossy sunscreen skin gel
mediating	reconciliation negotiate case conciliation
keeping	hoping bring wiping could some would
lithographs	drawings Picasso Dali sculptures Gauguin
pathogens	toxins bacteria organisms bacterial parasite
senses	grasp psyche truly clumsy naive innate

IDEA: PRF and CBD

- **Query Terms Expansion**
  - Consider a FIRE topic (query)
  - Obtain the initial set of documents and top k documents are assumed to be "*relevant*"
  - Apply CBD algorithm to get the initial set of directions in which search can be effectively carried out.
  - Choose terms representing the selected direction(s)
  - Query Terms weighting with entities

Pseudo Relevance Feedback (PRF)

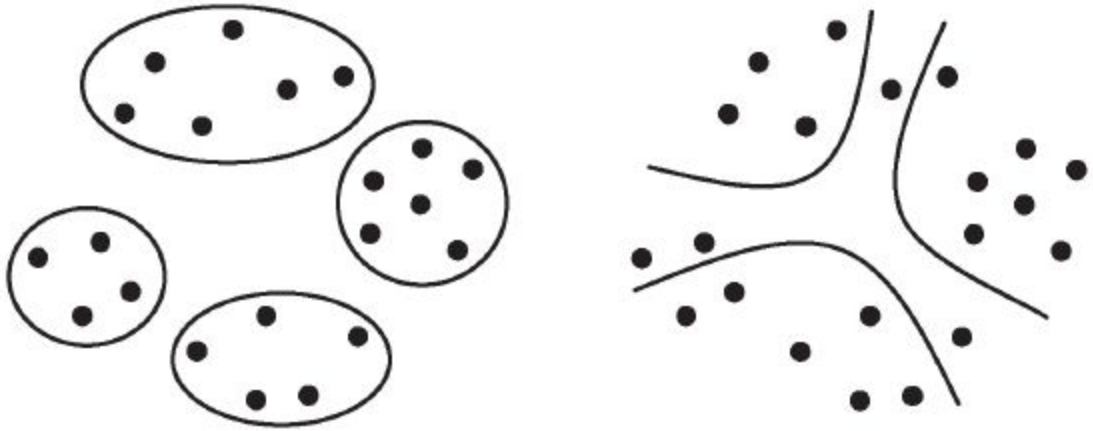
- In PRF, the expansion terms are based on co-occurrence relationships with query terms
  - Thus other terms which are lexically and semantically related are not explicitly captured
- Candidate (representative) set of Documents for the actual query (either title or desc)
  - Assume these documents as “PSEUDO” relevant documents
  - Represent terms of the pseudo relevant documents in the vector space
  - Represent the terms in tag cloud like arrangement

Tag Clouds – A Few Examples

- **Durga Puja** - pandal, puja, rules, durga, grandeur, calcutta, organisers, gimmicky, people, crowd
- **Howrah Bridge** - traffic, calcutta, bridge, howrah, port, setu, repairs, cpt, bankim, vehicles, structure, barge, trust, road, girders, police
- **Mamata Banerjee** - alliance, mamata, congress, trinamul, party, minister, bengal, cpm, left, meeting, calcutta
- **West bengal chief minister** - state, minister, party, bhattacharjee, bengal, west, chief, government, cpi

Clustering of Terms?

- Classical Clustering Vs. Clustering by Directions



- Neither cluster the search results nor the terms
- CLUE => direction that meets the search needs
- GUIDED NAVIGATION across the information pertaining to the specific user needs

Tag Cloud Based Approach for QE

- Consider the Query: **C**
- Search this query usin

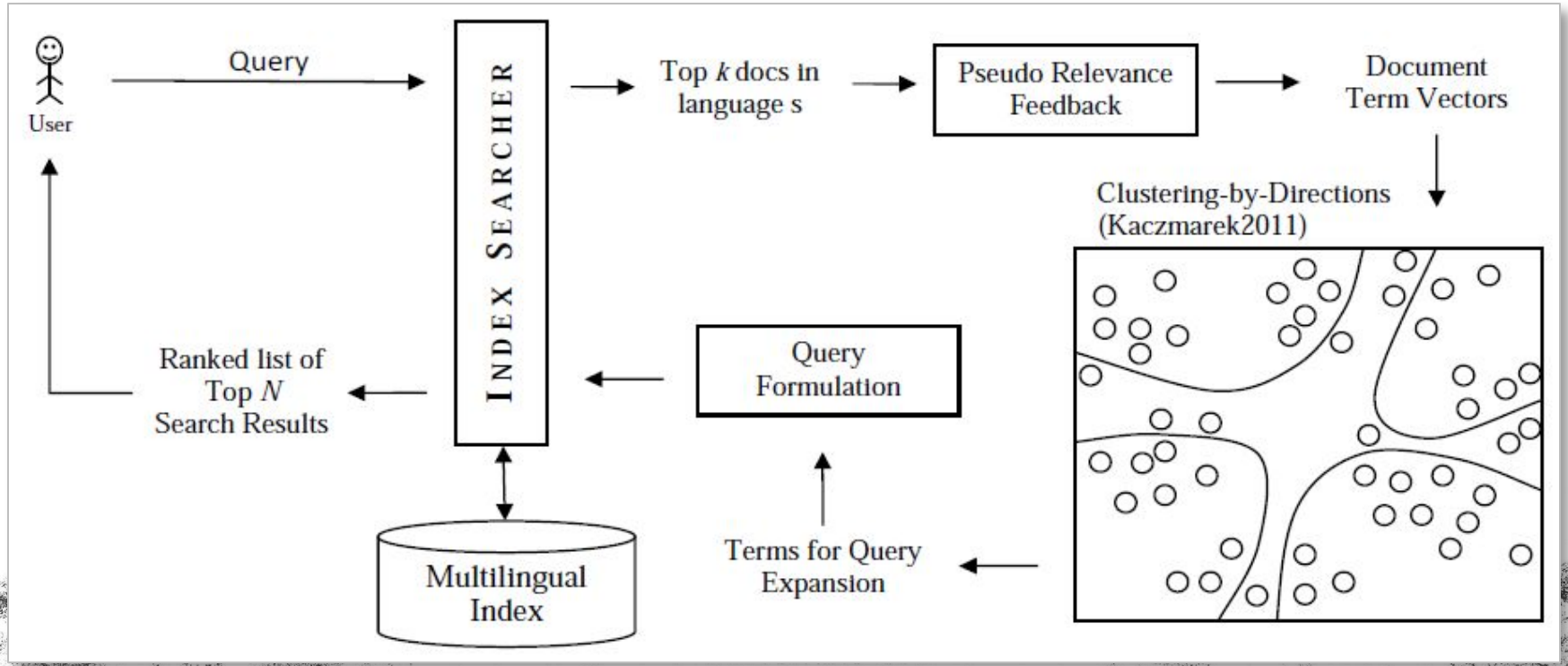
  "The space of knowledge"
- Clustering-By-Directions
- Idea here is to design an non-interactive algorithm that assists users to express their information needs

Clustering By Directions

Basic Steps (Adam 2011):

- calculate vectors which represent documents and distances between these vectors;
- select different directions;
- assign documents to directions and select terms which represent directions;
- select top k terms as candidate terms for query expansion

Proposed IR system - Architecture



Ranking Function – BM25 in Lucene

Given: a query Q containing $q_1, q_2, \ldots, q_n$

- The BM25 score of a document D *as follows:*

$$score(Q, D) = \sum_{i}^{n} idf(q_i) \cdot \frac{tf(q_i, D) \cdot (k_1 + 1)}{tf(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdoclength}})}$$

Where $tf(q_i, D)$ is the term frequency of q_i in the document D; $|D|$ is the length of the document D and $avgdoclength$ is the average document length in the text collection; $k_1, k_2 \in \{1.2, 2.0\}$ and $b = 0.85$ are parameters

- IDF (q_i) is computed as:

$$idf(q_i) = \log \frac{N - df(q_i) + 0.5}{df(q_i) + 0.5}$$

Where N is the total number of documents and $df(qi)$ is the number of documents containing the term q_i

Experiments

- Adhoc Track:
  - Corpus: News Documents Collection
  - Languages: Bengali, Hindi and English
- Resources Used:
  - Stop words
    - English: SMART stop words list
    - Other Languages: Resource from CLIA project
  - Stemming
    - English: Porter Stemmer
    - Others: CLIA stemmers`

Corpus Statistics and Topics

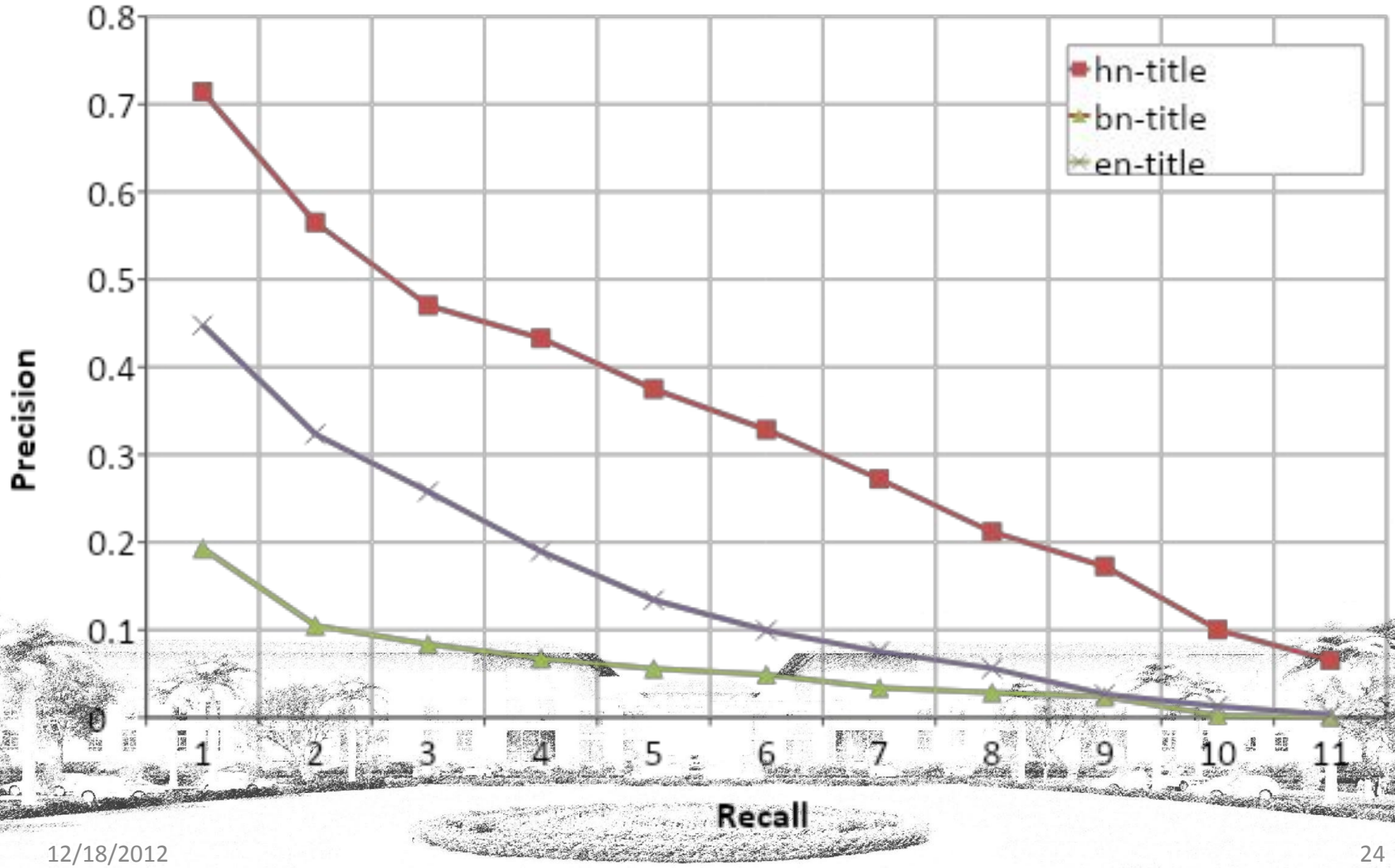
- FIRE 2012 Adhoc data collection:

Lang	#docs	#terms	TopicsIDS
Bengali	500,122	2,497,978	176-225
Hindi	331,599	1,164,526	176-225
Tamil	194,483	1,078,746	176-225
English	392,577	1,427,986	176-225

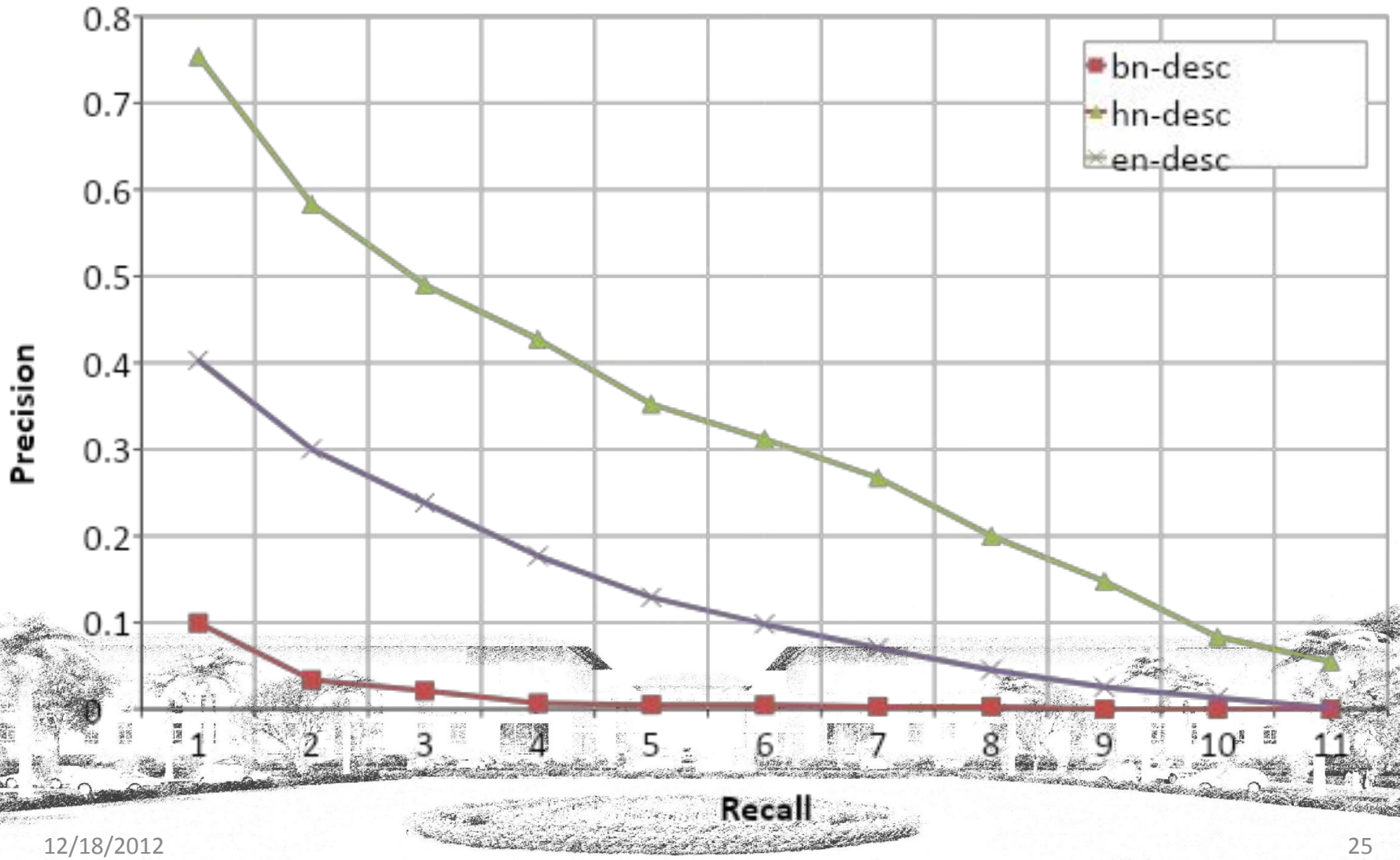
Topics (Queries)

- Range: 176 – 225 in English, Hindi, Bengali, Tamil, Gujarati, Telugu, Marathi, Odia, Punjabi and Assamese
- Fields in the topics:
  - ID – The Unique Query ID
  - TITLE – The Actual Query
  - DESC – The description of the query (Query Explained)
  - NARR – Narration about the Information need. This is not used for retrieval

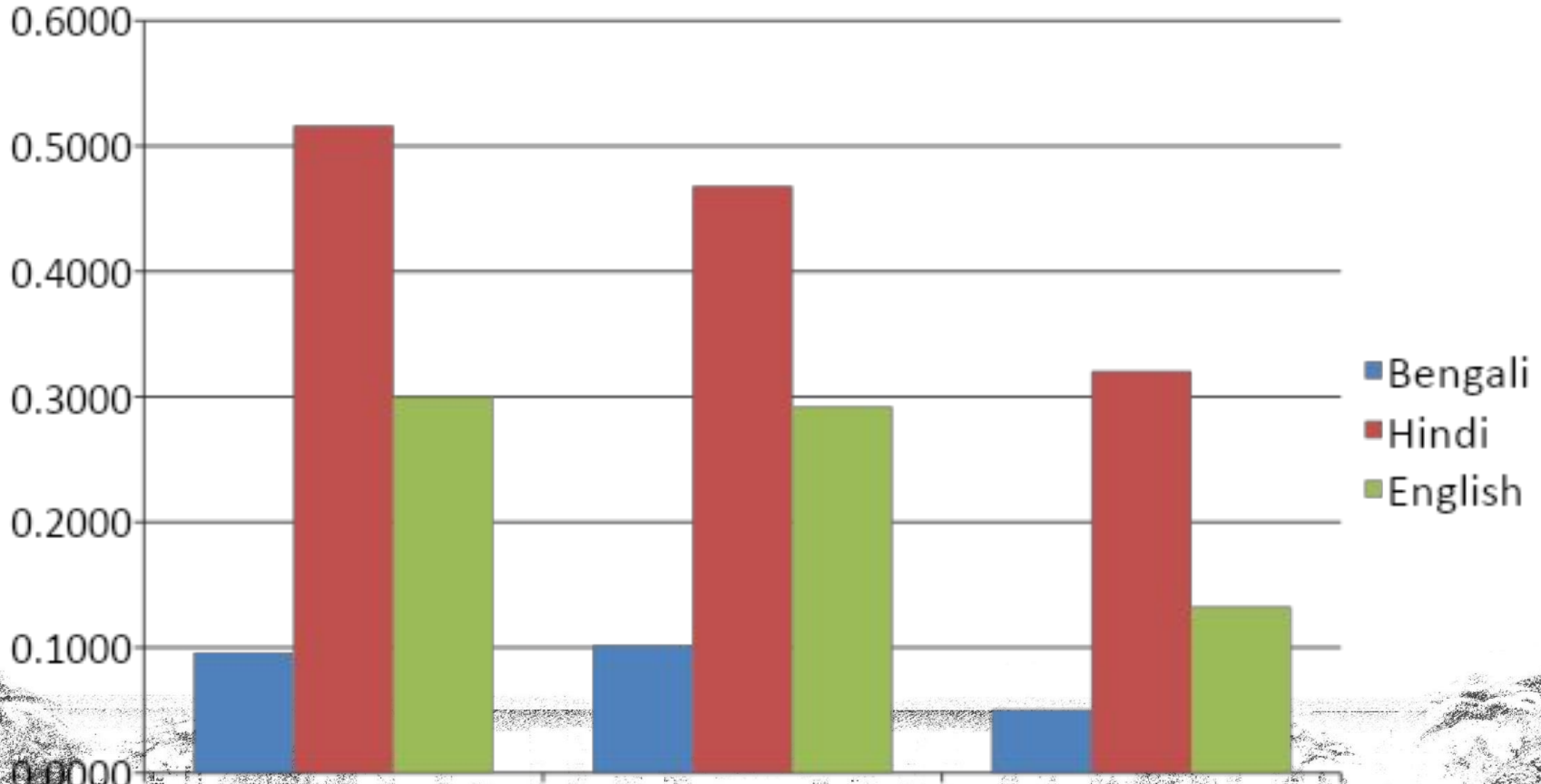
P-R curve: title used as query



P-R curve: desc used as query



Overall Performance



Observations:

- Issues (Query DRIFT): YSR Reddy death (results related to politics), MGNREGA scheme (expansion problems), Chamunda Temple stampede (Dharshan related terms), Adarsh Housing Society scam resignation (results are not related to Chief Minister Ashok Chavan's resignation), 2001 census India (not related to sex ratio / religion)
- Better Performance for Queries: Countries adopting EURO (180), Jaswant Singh BJP sacking (190), Prophet Muhammad cartoon agitation (199), NatWest Series 2002 result (200), Terrorist attack Indian Parliament (204), Polio eradication mission (205), Harbhajan Singh slapping Sreesanth (214), Imran Khan cancer hospital Pakistan (223), Satanic Verses controversy (225)

Observations:

- PRF-CBD approach is able to present terms that capture the variety of news items on this topic
- In Hindi mono retrieval with title, several queries, achieved a mean average precision greater than 0.7
- 60% of topics achieved map value of 0.5 or greater
- In English Mono Retrieval, almost 12 queries achieved map value of 0.25 and above with title
- No relevant docs @ Top 10: hi (6) en(20), bn (18)
- Query wise detailed error analysis are in half way

Any Impovements?

- ✔ Query Terms Expansion using PRF based CBD approach in monolingual documents retrieval

What's Next?

- ❑ PRF-CBD for Cross Lingual Documents Retrieval
- ❑ Expanding the queries in terms of entities and key phrases using unsupervised approaches to identify entities and key phrases in user queries
- ❑ Selecting contextual terms in presence of multiple translation / transliteration in CLIR

Query expansion at search engines

- Main source of query expansion at search engines: query logs
- **Example 1:**
  After issuing the query [herbs], users frequently search for [herbal remedies].
  - "herbal remedies" is potential expansion of "herb"
- **Example 2:**
  a) Users searching for [flower pix] frequently click on: photobucket.com/flower
  b) Users searching for [flower clipart] frequently click on the same URL.
  - "flower clipart" and "flower pix" are potential expansions of each other.

Summary

In this class, we focused on:

Query Expansion

i. Thesaurus based Approach
ii. Co-occurrence Based Approach
iii. Clustering By Directions

Reference

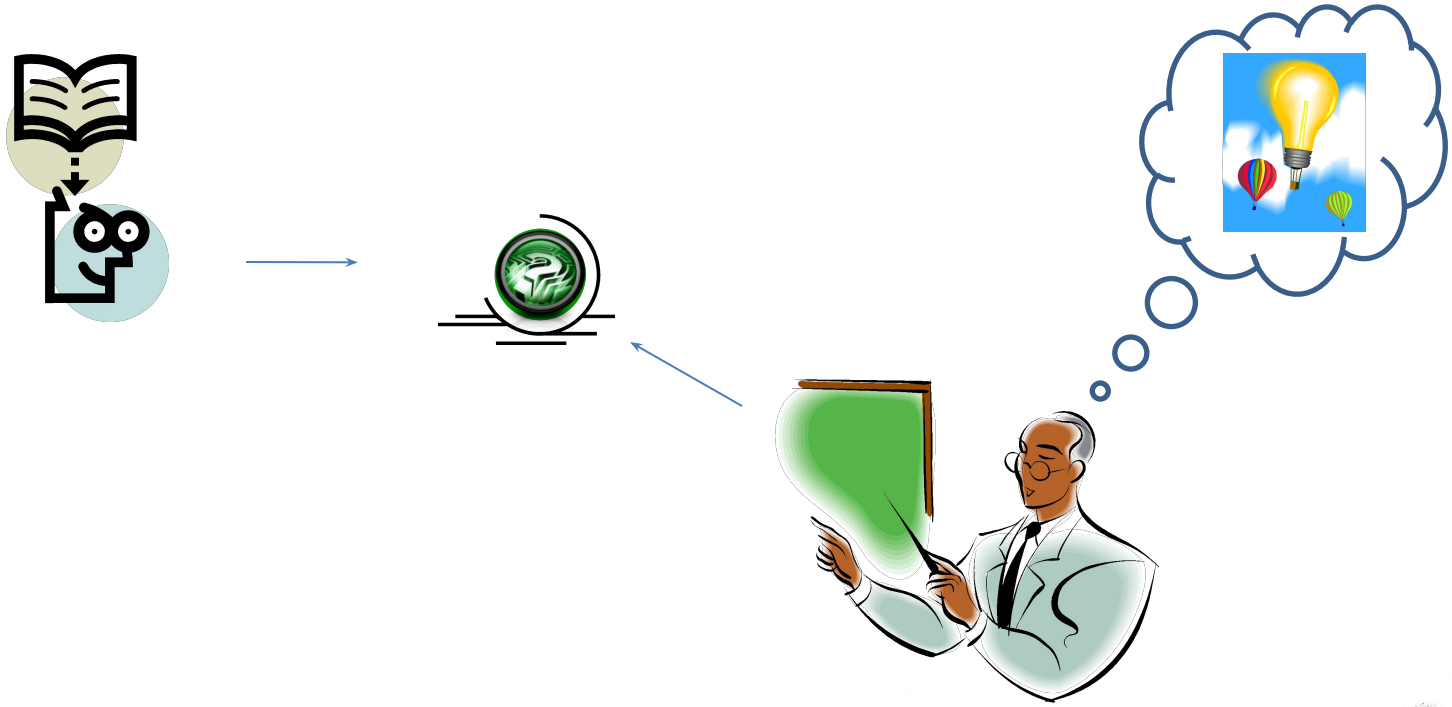
- **A. Kaczmarek. Interactive query expansion with the use of clustering-by-directions algorithm. Industrial Electronics, IEEE Transactions on, 58(8)(2011): 3168 –3173**
- A. Singhal, et al., Document Length Normalization. Inf. Process. Manage. 32(5) (1996): 619-633
- C. D. Manning, et al., Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA, 2008
- **S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond, Found. Trends Inf. Retr., 3(4)(2009): 333–389**
- S. E. Robertson & S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. SIGIR 1994: 232–241
- H. Chim and X. Deng, Efficient phrase-based document similarity for clustering, IEEE Trans. Knowl. Data Eng., 20(9) (2008): 1217–1229
- Manoj et al. Multilingual PRF: english lends a helping hand. SIGIR 2010: 659-666

Acknowledgements

Thanks to all IR Researchers:

1. Introduction to Information Retrieval Manning, Raghavan and Schutze, Cambridge University Press, 2008.
2. Search Engines Information Retrieval in Practice W. Bruce Croft, D. Metzler, T. Strohman, Pearson, 2009.
3. Information Retrieval Implementing and Evaluating Search Engines Stefan Büttcher, Charles L. A. Clarke and Gordon V. Cormack, MIT Press, 2010.
4. Modern Information Retrieval Baeza-Yates and Ribeiro-Neto, Addison Wesley, 1999.
5. Many Authors who contributed to SIGIR / WWW / KDD / ECIR / CIKM / WSDM and other top tier conferences
6. Prof. Mandar Mitra, Indian Statistical Institute, Kolkatata (https://www.isical.ac.in/~mandar/) for sharing the IR Evaluation Slides

Thanks …



… Questions ???