Evaluation in Information Retrieval

Mandar Mitra

Indian Statistical Institute

Outline

1 Preliminaries

2 Forums

3 Tasks

- Task 1: Morpheme extraction
- Task 2: RISOT
- Task 3: SMS-based FAQ retrieval
- Task 4: Microblog retrieval

Motivation

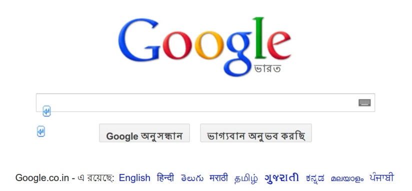
- Which is better: Heap sort or Bubble sort?

Motivation

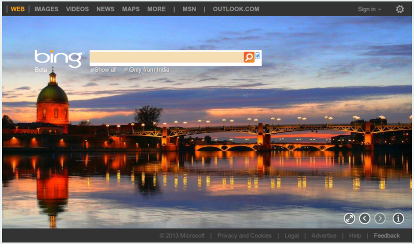
- Which is better: Heap sort or Bubble sort?

vs.

Which is better?


Google
ভারত
Google অনুসন্ধান
ভাগ্যবান অনুভব করছি
Google.co.in - এ রয়েছে: English हिन्दी తెలుగు मराठी தமிழ் ગુજરાતી ಕನ್ನಡ മലയാളം ਪੰਜਾਬੀ

or


WEB | IMAGES VIDEOS NEWS MAPS MORE | MSN | OUTLOOK.COM
Sign in
bing
Beta
Show all
Only from India
© 2013 Microsoft | Privacy and Cookies | Legal | Advertise | Help | Feedback

Motivation

- IR is an *empirical* discipline.

Motivation

- IR is an *empirical* discipline.
- Intuition can be wrong!
    - "sophisticated" techniques need not be the best
      e.g. rule-based stemming vs. statistical stemming

Motivation

- IR is an *empirical* discipline.
- Intuition can be wrong!
    - “sophisticated” techniques need not be the best
      e.g. rule-based stemming vs. statistical stemming
- Proposed techniques need to be validated and compared to existing techniques.

Cranfield method (Cleverdon et al., 60s)

Benchmark data

- Document collection
- Query / topic collection
- Relevance judgments - information about which document is relevant to which query

Cranfield method (Cleverdon et al., 60s)

Benchmark data

- Document collection
- Query / topic collection
- Relevance judgments - information about which document is relevant to which query

syllabus

question paper

correct answers

Cranfield method (Cleverdon et al., 60s)

Benchmark data

- Document collection

syllabus

- Query / topic collection

question paper

- Relevance judgments - information about which document is relevant to which query

correct answers

Assumptions

- relevance of a document to a query is objectively discernible
- all relevant documents contribute equally to the performance measures
- relevance of a document is independent of the relevance of other documents

Evaluation metrics

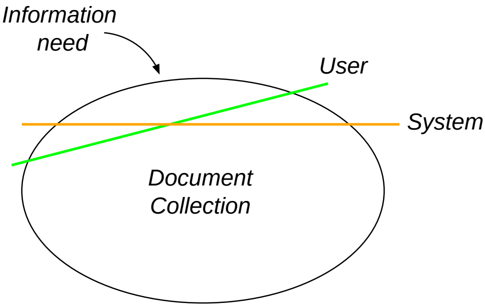
Background

- User has an information need.
- Information need is converted into a **query**.
- Documents are **relevant** or **non-relevant**.
- Ideal system retrieves all and only the relevant documents.

Evaluation metrics

Background

- User has an information need.
- Information need is converted into a **query**.
- Documents are **relevant** or **non-relevant**.
- Ideal system retrieves all and only the relevant documents.



Set-based metrics

$$\textbf{Recall} = \frac{\#(\text{relevant retrieved})}{\#(\text{relevant})} = \frac{\#(\text{true positives})}{\#(\text{true positives + false negatives})}$$

$$\textbf{Precision} = \frac{\#(\text{relevant retrieved})}{\#(\text{retrieved})} = \frac{\#(\text{true positives})}{\#(\text{true positives + false positives})}$$

$$\textbf{F} = \frac{1}{\alpha/P + (1-\alpha)/R} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

Metrics for ranked results

(Non-interpolated) average precision

Which is better?

1. Non-relevant
2. Non-relevant
3. Non-relevant
4. Relevant
5. Relevant

1. Relevant
2. Relevant
3. Non-relevant
4. Non-relevant
5. Non-relevant

Metrics for ranked results

(Non-interpolated) average precision

Rank	Type	Recall	Precision
1	Relevant	0.2	1.00
2	Non-relevant		
3	Relevant	0.4	0.67
4	Non-relevant		
5	Non-relevant		
6	Relevant	0.6	0.50

Metrics for ranked results

(Non-interpolated) average precision

Rank	Type	Recall	Precision
1	Relevant	0.2	1.00
2	Non-relevant		
3	Relevant	0.4	0.67
4	Non-relevant		
5	Non-relevant		
6	Relevant	0.6	0.50
∞	Relevant	0.8	0.00
∞	Relevant	1.0	0.00

Metrics for ranked results

(Non-interpolated) average precision

Rank	Type	Recall	Precision
1	Relevant	0.2	1.00
2	Non-relevant		
3	Relevant	0.4	0.67
4	Non-relevant		
5	Non-relevant		
6	Relevant	0.6	0.50
∞	Relevant	0.8	0.00
∞	Relevant	1.0	0.00

$$AvgP = \frac{1}{5}(1 + \frac{2}{3} + \frac{3}{6})$$

(5 relevant docs. in all)

Metrics for ranked results

(Non-interpolated) average precision

Rank	Type	Recall	Precision
1	Relevant	0.2	1.00
2	Non-relevant		
3	Relevant	0.4	0.67
4	Non-relevant		
5	Non-relevant		
6	Relevant	0.6	0.50
∞	Relevant	0.8	0.00
∞	Relevant	1.0	0.00

$$AvgP = \frac{1}{5}(1 + \frac{2}{3} + \frac{3}{6})$$

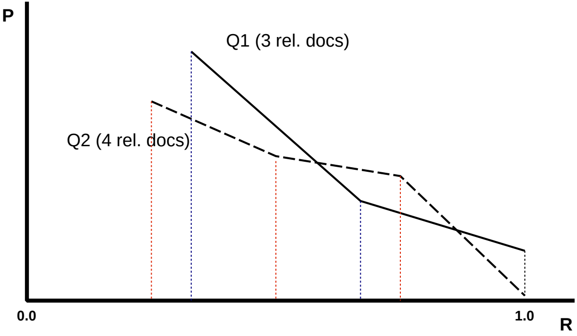
(5 relevant docs. in all)

$$\boxed{AvgP = \frac{1}{N_{Rel}} \sum_{d_i \in Rel} \frac{i}{Rank(d_i)}}$$

Metrics for ranked results

Interpolated average precision at a given recall point

- Recall points correspond to $\frac{1}{N_{Rel}}$
- N_{Rel} different for different queries



- Interpolation required to compute averages across queries

Metrics for ranked results

Interpolated average precision

$$\boxed{P_{int}(r) = \max_{r' \geq r} P(r')}$$

Metrics for ranked results

Interpolated average precision

$$\boxed{P_{int}(r) = \max_{r' \geq r} P(r')}$$

11-pt interpolated average precision

Rank	Type	Recall	Precision
1	Relevant	0.2	1.00
2	Non-relevant		
3	Relevant	0.4	0.67
4	Non-relevant		
5	Non-relevant		
6	Relevant	0.6	0.50
∞	Relevant	0.8	0.00
∞	Relevant	1.0	0.00

Metrics for ranked results

Interpolated average precision

$$\boxed{P_{int}(r) = \max_{r' \geq r} P(r')}$$

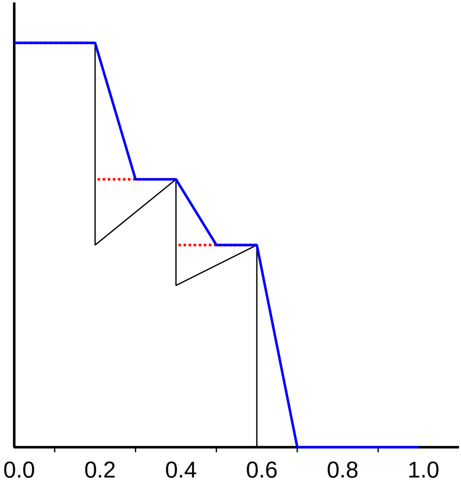
11-pt interpolated average precision

Rank	Type	Recall	Precision
1	Relevant	0.2	1.00
2	Non-relevant		
3	Relevant	0.4	0.67
4	Non-relevant		
5	Non-relevant		
6	Relevant	0.6	0.50
∞	Relevant	0.8	0.00
∞	Relevant	1.0	0.00

R	Interp. P
0.0	1.00
0.1	1.00
0.2	1.00
0.3	0.67
0.4	0.67
0.5	0.50
0.6	0.50
0.7	0.00
0.8	0.00
0.9	0.00

Metrics for ranked results

11-pt interpolated average precision



Metrics for sub-document retrieval

Let p_r - document part retrieved at rank r

$rsize(p_r)$ - amount of relevant text contained by p_r

$size(p_r)$ - total number of characters contained by p_r

T_{rel} - total amount of relevant text for a given topic

$$\begin{aligned} P[r] &= \frac{\sum_{i=1}^{r} rsize(p_i)}{\sum_{i=1}^{r} size(p_i)} \\ R[r] &= \frac{1}{T_{rel}} \sum_{i=1}^{r} rsize(p_i) \end{aligned}$$

Metrics for ranked results

- **Precision at k (P@k)** - precision after **k** documents have been retrieved
  - easy to interpret
  - not very stable / discriminatory
  - does not average well
- **R precision** - precision after N_{Rel} documents have been retrieved

Cumulated Gain

Idea:

- Highly relevant documents are more valuable than marginally relevant documents
- Documents ranked low are less valuable

Cumulated Gain

Idea:

- Highly relevant documents are more valuable than marginally relevant documents
- Documents ranked low are less valuable

$$Gain \in \{0, 1, 2, 3\}$$

$$G = \langle 3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \ldots \rangle$$

$$CG[i] = \sum_{j=1}^{i} G[i]$$

(n)DCG

$$DCG[i] = \begin{cases} CG[i] & \text{if } i < b \\ DCG[i-1] + G[i]/\log_b i & \text{if } i \geq b \end{cases}$$

(n)DCG

$$DCG[i] = \begin{cases} CG[i] & \text{if } i < b \\ DCG[i-1] + G[i]/\log_b i & \text{if } i \geq b \end{cases}$$

$$\textbf{Ideal } G = \langle 3, 3, \ldots, 3, 2, \ldots, 2, 1, \ldots, 1, 0, \ldots \rangle$$

$$nDCG[i] = \frac{DCG[i]}{\textbf{Ideal } DCG[i]}$$

Mean Reciprocal Rank

- Useful for *known-item* searches with a single target
- Let r_i — rank at which the “answer” for query i is retrieved.
  Then reciprocal rank = $1/r_i$

  Mean reciprocal rank (MRR) = $\sum_{i=1}^{n} \frac{1}{r_i}$

Assumptions

- All relevant documents contribute equally to the performance measures.
- Relevance of a document to a query is objectively discernible.
- Relevance of a document is independent of the relevance of other documents.

Assumptions

- All relevant documents contribute equally to the performance measures.
- Relevance of a document to a query is objectively discernible.
- Relevance of a document is independent of the relevance of other documents.
- All relevant documents in the collection are known.

Assessor agreement

- Judges / assessors may not agree about relevance.

Example (MANNING ET AL.)

	Yes$_1$	No$_1$	Total$_2$
Yes$_2$	300	20	320
No$_2$	10	70	80
Total$_1$	310	90	400

$$P(A) = (300 + 70)/400 = 370/400 = 0.925$$
$$P(\text{nrel}) = (80 + 90)/(400 + 400) = 0.2125$$
$$P(\text{rel}) = (320 + 310)/(400 + 400) = 0.7878$$
$$P(E) = P(\text{non-rel})^2 + P(\text{rel})^2 = 0.665$$
$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} = \frac{0.925 - 0.665}{1 - 0.665} = 0.776$$

- Rules of thumb:
  $\kappa > 0.8$ — good agreement
  $0.67 \leq \kappa \leq 0.8$ — fair agreement
  $\kappa < 0.67$ — poor agreement

Pooling

- Exhaustive relevance judgments may be infeasible.
- Pool top results obtained by various systems and assess the pool.
- *Unjudged documents are assumed to be non-relevant.*

Pooling

- Exhaustive relevance judgments may be infeasible.
- Pool top results obtained by various systems and assess the pool.
- *Unjudged documents are assumed to be non-relevant.*
- A wide variety of models, retrieval algorithms is important.
- **Manual interactive retrieval** is a must.

Pooling

- Exhaustive relevance judgments may be infeasible.
- Pool top results obtained by various systems and assess the pool.
- *Unjudged documents are assumed to be non-relevant.*
- A wide variety of models, retrieval algorithms is important.
- **Manual interactive retrieval** is a must.

Can unbiased, incomplete relevance judgments be used to reliably compare the relative effectiveness of different retrieval strategies?

Bpref

- Based on number of times judged nonrelevant documents are retrieved before relevant documents

  Let R - set of relevant documents for a topic
  N - set of first $|R|$ judged non-rel docs retrieved

$$bpref = \frac{1}{|R|} \sum_{r \in R} (1 - \frac{|n \text{ ranked higher than } r|}{|R|})$$

$$bpref10 = \frac{1}{|R|} \sum_{r \in R} (1 - \frac{|n \text{ ranked higher than } r|}{10 + |R|})$$

Bpref

- With complete judgments:
  system rankings generated based on MAP and bpref10 are highly correlated
- When judgments are incomplete:
  system rankings generated based on bpref10 are more stable

Outline

1 Preliminaries

2 Forums

3 Tasks

- Task 1: Morpheme extraction
- Task 2: RISOT
- Task 3: SMS-based FAQ retrieval
- Task 4: Microblog retrieval

TREC

`http://trec.nist.gov`

- Organized by NIST every year since 1992
- Typical tasks
  - adhoc
    - user enters a search topic for a one-time information need
    - document collection is static
  - routing/filtering
    - user's information need is persistent
    - document collection is a stream of incoming documents
  - question answering

TREC data

- Documents
    - Genres:
        - news (AP, LA Times, WSJ, SJMN, Financial Times, FBIS)
        - govt. documents (Federal Register, Congressional Records)
        - technical articles (Ziff Davis, DOE abstracts)
    - Size: 0.8 million documents – 1.7 million web pages (cf. Google indexes several billion pages)
- Topics
    - title
    - description
    - narrative

Enterprise track

- Goal: work with enterprise data (intranet pages, email archives, document repositories)
- Corpus: crawled from W3C
  - lists.w3.org: $\sim$ 200,000 docs, $\sim$ 2 GB
  - documents are html-ised archives of mailing lists (email header information recoverable)

Enterprise track tasks

Known item search

- Scenario: user searches for a particular message that is known to exist
- Test data: topic + corresponding (unique) message
- Measures: mean reciprocal rank (MRR), success at 10 docs.
- Results: best groups obtained MRR $\approx$ 0.6, S@10 $\approx$ 0.8

Enterprise track tasks

Discussion search

- Scenario: user searches for an argument / discussion on an issue
- Test data: topic / issue + relevant messag
  endframe
  - irrelevant
  - partially relevant (does not take a stand)
  - relevant (takes a pro/con stand)
- Measures: MAP, R-precision, etc.
- Results: “strict” and “lenient” evaluations were strongly correlated

Enterprise track tasks

Expert search

- Scenario: user searches for names of persons who are experts on a specified topic
- Test data: working groups of the W3C, members of the working groups
- Measures: MAP, R-precision, etc.

Enterprise track

Hiccups

- Discussion search: no dry runs
- Some topics more amenable than others to a pro/con discussion
- Relevance judgments do not include orientation information
- Assessor agreement: ranking done on the basis of *primary* and *secondary* assessors correlated, but not “essentially identical”

CLEF

`http://www.clef-campaign.org/`

- CLIR track at TREC-6 (1997), CLEF started in 2000
- Objectives:
  - to provide an infrastructure for the testing and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts
  - to construct test-suites of reusable data that can be employed by system developers for benchmarking purposes
  - to create an R&D community in the cross-language information retrieval (CLIR) sector

CLEF tasks

- Monolingual retrieval
- Bilingual retrieval
  - queries in language X
  - document collection in language Y
- Multi-lingual retrieval
  - queries in language X
  - multilingual collection of documents (e.g. English, French, German, Italian)
  - results include documents from various collections and languages in a single list
- Other tasks: spoken document retrieval, image retrieval

NTCIR

`http://research.nii.ac.jp/ntcir`

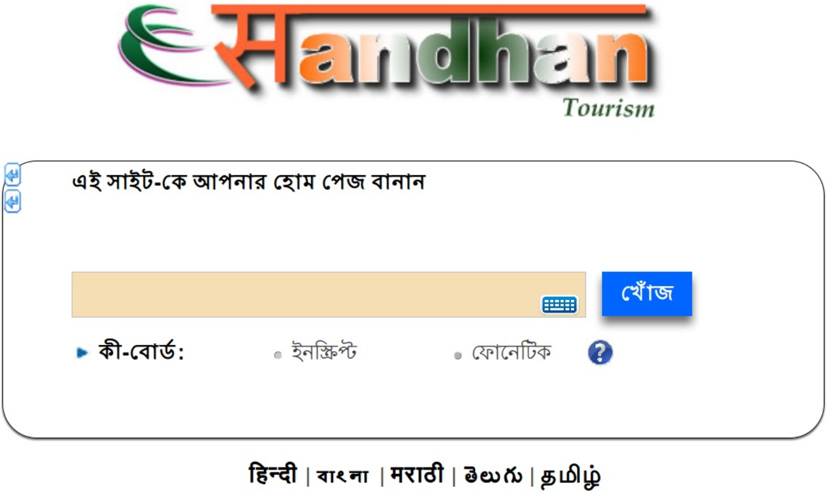
- Started in late 1997
- Held every 1.5 years at NII, Japan
- Focus on East Asian languages (Chinese, Japanese, Korean)
- Tasks
  - cross-lingual retrieval
  - patent retrieval
  - geographic IR
  - opinion analysis

FIRE

- Forum for Information Retrieval Evaluation
  `http://www.isical.ac.in/~fire`
- Evaluation component of a DIT-sponsored, consortium mode project
- Assigned task: create a portal where
  1. a user will be able to give a query in one Indian language;
  2. s/he will be able to access documents available in the language of the query, Hindi (if the query language is not Hindi), and English,
  3. all presented to the user in the language of the query.
- Languages: Bangla, Hindi, Marathi, Punjabi, Tamil, Telugu

Sandhan: a search engine

`http://tdil-dc.in/sandhan`



सandhan Tourism

এই সাইট-কে আপনার হোম পেজ বানান

খোঁজ

কী-বোর্ড: ○ ইনস্ক্রিপ্ট ○ ফোনেটিক ?

हिन्दी | বাংলা | मराठी | తెలుగు | தமிழ்

FIRE: goals

- To encourage research in South Asian language Information Access technologies by providing reusable large-scale test collections for ILIR experiments
- To provide a common evaluation infrastructure for comparing the performance of different IR systems
- To explore new Information Retrieval / Access tasks that arise as our information needs evolve, and new needs emerge
- To investigate evaluation methods for Information Access techniques and methods for constructing a reusable large-scale data set for ILIR experiments.
- To build language resources for IR and related language processing tasks

FIRE: tasks

- Ad-hoc monolingual retrieval
  - Bengali, Hindi Marathi and English
- Ad-hoc cross-lingual document retrieval
  - documents in Bengali, Hindi, Marathi, and English
  - queries in Bengali, Hindi, Marathi, Tamil, Telugu , Gujarati and English
  - Roman transliterations of Bengali and Hindi topics
- CL!NSS: Cross-Language !ndian News Story Search
- MET: Morpheme Extraction Task (MET)
- RISOT: Retrieval from Indic Script OCR'd Text
- SMS-based FAQ Retrieval
- Older tracks:
  - Retrieval and classification from mailing lists and forums
  - Ad-hoc Wikipedia-entity retrieval from news documents

FIRE: datasets

Documents

- Bengali: Anandabazar Patrika (123,047 docs)
- Hindi: Dainik Jagran (95,215 docs) +
  Amar Ujala (54,266 docs)
- Marathi: Maharashtra Times, Sakal (99,275 docs)
- English: Telegraph (125,586 docs)

- All from the Sep 2004 - Sep 2007 period
- All content converted to UTF-8
- Minimal markup

FIRE: datasets

Topics

- 225 topics
- Queries formulated parallely in Bengali, Hindi by browsing the corpus
- Refined based on initial retrieval results
  - ensure minimum number of relevant documents per query
  - balance easy, medium and hard queries
- Translated into Marathi, Tamil, Telugu , Gujarati and English
- TREC format (title + desc + narr)

FIRE: topics

Example:

```
<title> Nobel theft

<desc>
Rabindranath Tagore's Nobel Prize medal was stolen from Santiniketan. The
document should contain information about this theft.

<narr>
A relevant document should contain information regarding the
missing Nobel Prize Medal that was stolen along with some other
artefacts and paintings on 25th March, 2004. Documents containing
reports related to investigations by government agencies like CBI
/ CID are also relevant, as are articles that describe public
reaction and expressions of outrage by various political parties.
```

Participants

AU-KBC

IIT Bombay

Jadavpur U.

IBM

Microsoft Research

Dublin City U.

U. of Maryland

U. Neuchatel, Switzerland

U. North Texas

U. Tampere

Outline

1 Preliminaries

2 Forums

3 Tasks
- Task 1: Morpheme extraction
- Task 2: RISOT
- Task 3: SMS-based FAQ retrieval
- Task 4: Microblog retrieval

Task overview

Morpheme: smallest meaningful units of language

Task overview

Morpheme: smallest meaningful units of language

- Objective: discover morphemes in (morphologically rich) Indian languages
- Started in 2012
- Offered in Bengali, Gujarati, Hindi, Marathi, Odia, Tamil

Details

- Participants need to submit a working program (morpheme extraction system)
- Specifications:
  Input: large lexicon (provided as test data to the participants)
  Output: two column file containing list of words + morphemes, separated by tab
- Evaluation: use system output for *stemming* within an IR system
  - System: TERRIER (http://www.terrier.org)
  - Task: FIRE 2011 adhoc task
  - Metric: MAP

Results

Bengali

Team	MAP
Baseline	0.2740
CVPR-Team1	0.3159
DCU	0.3300
IIT-KGP	0.3225
ISM	0.3013
JU	**0.3307**

Task overview

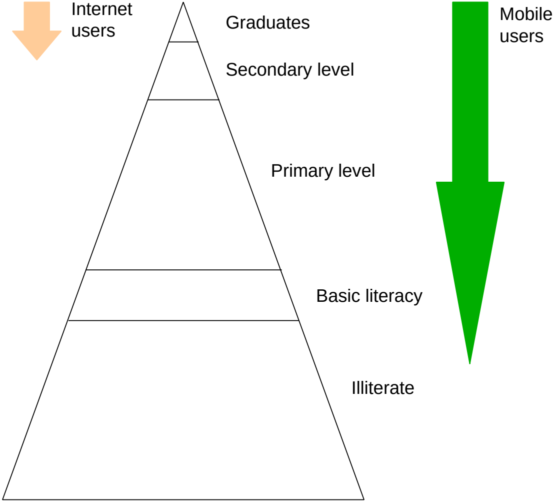
RISOT: retrieval of Indic script OCR'd text

- Data
  - 62,875 Bengali newspaper articles from FIRE 2008/2010 collection
    + 66 topics
  - 94,432 Hindi newspaper articles
    + 28 topics
- Documents automatically rendered and OCR'd

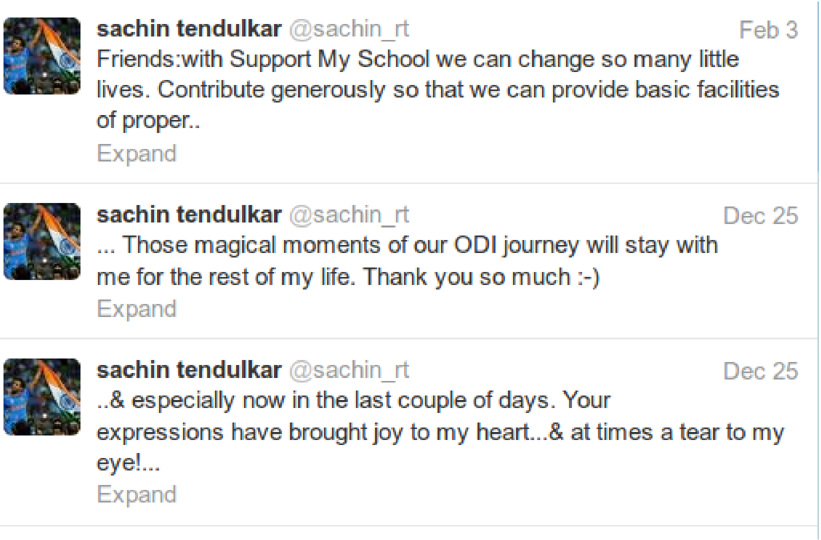
Results

Source	MAP
Clean text	0.2567
OCR'd text	0.1791
OCR'd text + processing	0.1974

Motivation


Internet users
Graduates
Secondary level
Primary level
Basic literacy
Illiterate
Mobile users

Background



sachin tendulkar @sachin_rt Feb 3

Friends:with Support My School we can change so many little lives. Contribute generously so that we can provide basic facilities of proper..

Expand

sachin tendulkar @sachin_rt Dec 25

... Those magical moments of our ODI journey will stay with me for the rest of my life. Thank you so much :-)

Expand

sachin tendulkar @sachin_rt Dec 25

..& especially now in the last couple of days. Your expressions have brought joy to my heart...& at times a tear to my eye!...

Expand

Challenges

Vocabulary mismatch

- Tweets are short (maximum of 140 characters each)
- Not always written maintaining formal grammar and proper spelling

Challenges

Vocabulary mismatch

- Tweets are short (maximum of 140 characters each)
- Not always written maintaining formal grammar and proper spelling

$\boxed{\Rightarrow \text{ Query keywords may not match a relevant tweet}}$

Query expansion

Definition. Add related words to query.

Example:

harmful effects of tobacco

Query expansion

Definition. Add related words to query.

Example:

harmful effects of tobacco

⇓

+ health, cancer, cigarette, smoking, gutkha

Relevance feedback

- Original query is used to retrieve some number of documents.
- User examines some of the retrieved documents and provides feedback about which documents are relevant and which are non-relevant.
- System uses the feedback to “learn” a better query:
  - select/emphasize words that occur more frequently in relevant documents than non-relevant documents;
  - eliminate/de-emphasize words that occur more frequently in non-relevant than in relevant documents.
- Resulting query should bring in more relevant documents and fewer non-relevant documents.

Relevance feedback

- Original query is used to retrieve some number of documents.
- User examines some of the retrieved documents and provides feedback about which documents are relevant and which are non-relevant.
- System uses the feedback to “learn” a better query:
  - select/emphasize words that occur more frequently in relevant documents than non-relevant documents;
  - eliminate/de-emphasize words that occur more frequently in non-relevant than in relevant documents.
- Resulting query should bring in more relevant documents and fewer non-relevant documents.

Blind/adhoc/pseudo relevance feedback: In the absence of feedback, *assume* top-ranked documents are relevant.

Query expansion for tweet retrieval

Bandyopadhyay et al., IJWS 2013

Query reformulation

- Initial query is used to retrieve d tweets
- *All distinct words* occurring in these tweets are used as new query

Query expansion for tweet retrieval

Bandyopadhyay et al., IJWS 2013

Query reformulation

- Initial query is used to retrieve d tweets
- *All distinct words* occurring in these tweets are used as new query

Collection enrichment: Use external source, e.g. Google

TREC 2011 microblog task

HTTP status	No. of tweets
Total crawled	**16,087,002**
301 (moved permanently)	987,866
302 (found but retweet)	1,054,459
403 (forbidden)	404,549
404 (not found)	458,388
Unknown[3]	3
200 (OK)	13,181,737
After filtering	**9,363,521**

Results

Run Name	P@30	MAP
B	0.3231	0.1938
PRF	0.3578	0.2283
	(+10.74%)	(+17.80%)
QR	0.3891	0.2515
	(+20.43%)	(+29.77%)
QR+PRF	0.4150	0.2754
	(+28.44%)	(+42.11%)
TGQR	0.4218	0.2824
	(+30.55%)	(+45.72%)
TGQE	0.4238	0.2819
	(+31.17%)	(+45.46%)

References

- *Introduction to Modern Information Retrieval.* Salton, McGill. McGraw Hill, 1983.
- *An Introduction to Information Retrieval.* Manning, Raghavan, Schutze.

  `http://www-csli.stanford.edu/~schuetze/information-retrieval-book.html`
- *Retrieval Evaluation with Incomplete Information.* Buckley, Voorhees. SIGIR 2004.
- `http://trec.nist.gov`
- *Cross-Language Evaluation Forum: Objectives, Results, Achievements.* Braschler, Peters. Information Retrieval, 7:12, 2004.
- `http://research.nii.ac.jp/ntcir`