

PART 2 Tutorial @ ECIR
Toulouse
6th Apr. 2009

Thomas Mandl
Information Science
University of Hildesheim
mandl@uni-hildesheim.de

Metrics for System Comparison

1

Evaluation research

- Alternative metrics are frequently being discussed
- Which metric measures which system behavior?
- Large amount of data is available now from evaluations
 - Validation is possible
 - Frequently strong correlation between metrics are observed

Mandl: Current Developments in Information Retrieval Evaluation

Metrics

- Requirements for Metrics
 - Correlation to user satisfaction
 - Easily interpretable
 - Robust
 - Work under many conditions
 - Capability for discrimination between systems
 - Discriminative power
 - „good“ systems should be evaluated better
 - Absolute values not very important

Mandl: Current Developments in Information Retrieval Evaluation 3

Mean Reciprocal Rank

- Reciprocal of the rank of the first relevant document returned
- Appropriate for known item search
- e.g.: first relevant doc ist at position 3

-> $MRR = 1/3$

Mandl: Current Developments in Information Retrieval Evaluation

Prec at N

- Precision after n documents
 - Frequently used
 - V.a. N=10
 - easily interpretable
 - Reasonable for Web Retrieval
 - very instable
 - Relies on few information about the system performance
 - All queries are evaluated on the same level
 - Position of relevant results is irrelevant

Mandl: Current Developments in Information Retrieval Evaluation

R-Prec

- Precision after r documents
- More stable measurus than Precision at n
- R = number of known relevant documents for the query

– Relies on less information than MAP

Mandl: Current Developments in Information Retrieval Evaluation

BPref

- Binary preference
- Problem:
 - Test collections increase strongly (concerning documents)
 - Percentaged there are less relevance decisions available
 - CLEF ad hoc: ca. 5% of the documents
 - TREC Web Track: ca. 0,4% of the documents
 - Not evaluated documents have influence on MAP
 - Count as not relevant
- Idea:
 - Only use evaluated documents

Bpref (Buckley & Voorhees 2004)

- Idea:
 - Not evaluated documents are „ignored“
 - How often do relevant documents occur in front of not relevant documents?
- Calculation
 - Generate pairs from all relevant and not relevant documents
 - How many of these pairs occur in a ranking in the „correct“ order?
- Bpref is already the standard measure at TREC

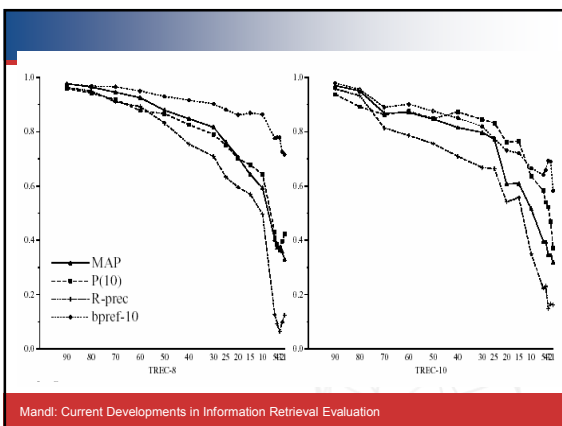
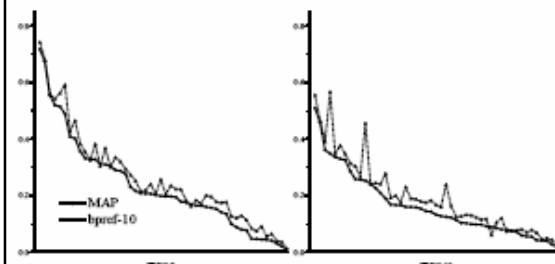
bpref

$$bpref = \frac{1}{R} \sum_r 1 - \frac{|n \text{ ranked higher than } r|}{R}$$

- bpref is very robust
- Problems occur for topics with only a few relevant documents

BPref and MAP

- Consistence per topic



Overview

- Cranfield Paradigm
 - Introduction
 - Validity
- Evaluation Metrics
 - Binary relevance
 - Multi level relevance
 - Evaluation Initiatives
- Topic Specific Analysis
 - Results
 - Optimization
- User Studies

Cumulative Gain (CG)

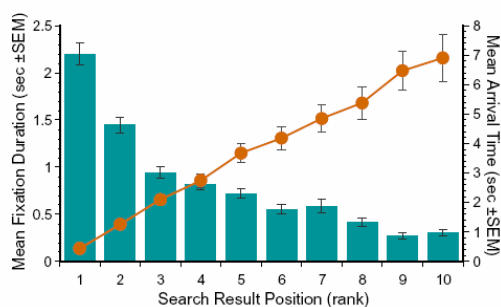
- The further a user reviews the list of result documents, the more “benefit” he/she gets
- What's the „benefit“, if he/she inspects up to x documents?
- To add the relevance until a position in the list
- Similar: comparison to an optimal ranking

CG

- Relevance values 0 – 3 are used:
 $G' = \langle 3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \dots \rangle$
- Cumulated Gain (CG)
 $CG' = \langle 3, 5, 8, 8, 8, 9, 11, 13, 16, 16, \dots \rangle$

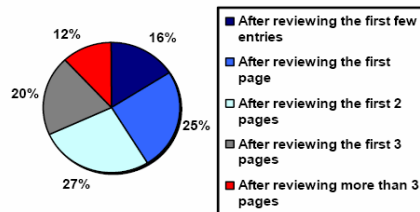
Järvelin & Kekäläinen 2002

Time per entry in hit list



How many result pages does a user inspect ?

"When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)"



DCG

- Discounted Cumulated Gain (DCG)
- do relevant documents occur further down in the list, the benefit is reduced
 - Division by the logarithm of the position in the hit list
 - Basis of the Logarithm determines the strength of the „Discount“
 - User model
 - often 2 or e is used

$G' = \langle 3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \dots \rangle$

$DG = \langle 3, 2, 1.89, 0, 0, 0.38, \dots \rangle$

• $DCG' = \langle 3, 6, 8.89, 8.89, 6.89, 7.28, 7.99, 8.66, \dots \rangle$

Järvelin & Kekäläinen 2002

Normalized Discounted Cumulated Gain (NDCG)

- Normalisation an der optimal sequence
 - Ideal sequence
 - Calculate its DCG
 - At each position the real DCG is divided by the ideal DCG

Ideal vector $I = \langle 3, 3, 3, 2, 2, 2, 1, 1, 1, 1, 0, 0, 0, \dots \rangle$

$CG_I = \langle 3, 6, 9, 11, 13, 15, 16, 17, 18, 19, 19, 19, 19, \dots \rangle$

$DCG_I = \langle 3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 11.21, \dots \rangle$

→ $NDCG' = \langle 1, 0.83, 0.89, 0.73, 0.62, 0.6, 0.69, 0.76, \dots \rangle$

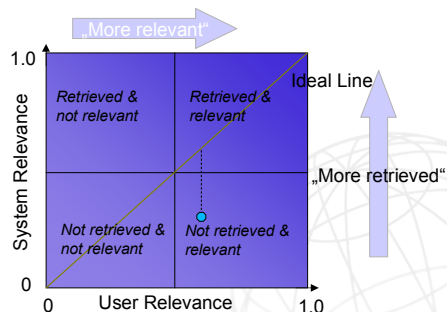
Average Distance Measure

- Continuous relevance values
- How large is the distance between the user and the system relevance score?

$$ADM_q = \frac{\sum_{d_j \in D} |SRS_q(d_j) - URS_q(d_j)|}{|D|}$$

(della Mea et al. 2006)

Mandl: Current Developments in Information Retrieval Evaluation



Mandl: Current Developments in Information Retrieval Evaluation

Metrics

- Many new metrics are being defined continuously
 - Frustration (Korfhage 1997)
 - Usefulness (Frei & Schäuble 1991)
 - Q-measure: based on Cumulative Gain, penalty for going down the list (Sakai 2004)
 - eXtended Cumulative Gain (XCG): relevance value functions modelling different user behavior (Kazai 2004)
 -

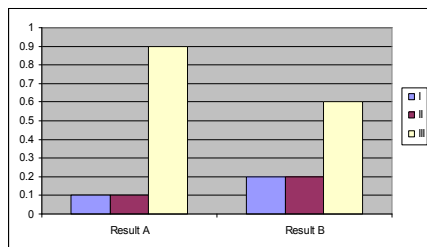
Mandl: Current Developments in Information Retrieval Evaluation

Aggregation of Several Queries

- Normally simple, arithmetic mean
- All individual „observations“ (topics) have the same contribution

Mandl: Current Developments in Information Retrieval Evaluation

Which system is better?



2 Systems, 3 Queries

Mandl: Current Developments in Information Retrieval Evaluation

23

„IR Psychology“

«The **unhappy customer**, on average, will tell **27** other people ...»

«**Dissatisfied customers** tell an average of **ten** other people about their bad experience. Twelve percent tell up to twenty people.»

→ **Bad news travels fast.**

Mandl: Current Developments in Information Retrieval Evaluation Credit to Jacques Savoy for slide 24


24

Robustness

capable of coping well with variations (sometimes unpredictable variations) in its operating environment with minimal damage, alteration or loss of functionality.

Mandl: Current Developments in Information Retrieval Evaluation

„IR Psychology“

On the other hand, satisfied customers will tell an average of *five* people about their positive experience. 

→ **Good news travels somewhat slower**

A Retrieval system should work robustly
Outlier on the lower end need to be avoided

Mandl: Current Developments in Information Retrieval Evaluation Credit to Jacques Savoy for slide 26

„IR Psychology“

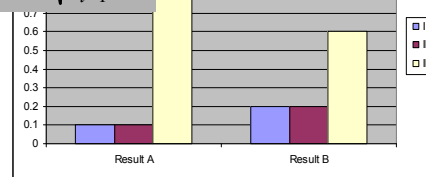
- Any system component can lead to dissatisfaction
- Very good results, bad interaction
– -> dissatisfied user
- Very good interaction, bad results
– -> dissatisfied user

Mandl: Current Developments in Information Retrieval Evaluation

27

GMAP

$$geoAve = \sqrt[n]{\prod_{i=1}^n x_i}$$



GeoAve	A	0.21	GeoAve	B	0.29
MAP	A	0.37	MAP	B	0.33

Mandl: Current Developments in Information Retrieval Evaluation

Robustness

- Robust performance over all queries instead of high average performance
- “Difficult” queries gain stronger weight
- But: what happens for Queries with MAP = 0 ?

Mandl: Current Developments in Information Retrieval Evaluation

29

Long-term aim

- „More work needs to be done on customizing methods for each topic“ (Harman 2005)
- Little Research
 - RIA Workshop
 - SIGIR Workshop on Topic difficulty
 - TREC Robust Track (until 2005)
 - ...

Mandl: Current Developments in Information Retrieval Evaluation

30

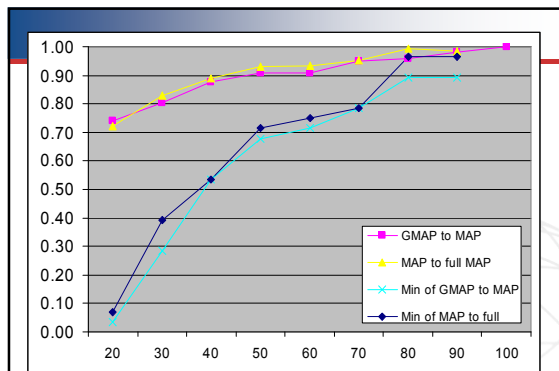
GMAP

- Alternative form of calculation:
 - Mean of logarithmic single measurements
 - Original measurements are transformed in a non-linear way

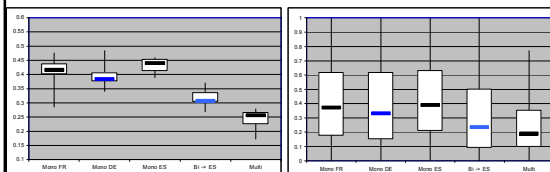
$$GMAP = \sqrt[n]{\prod_{i=1}^n x_i}$$

- Other transformations are possible (e.g. Log Basis)
- Can be based on user models

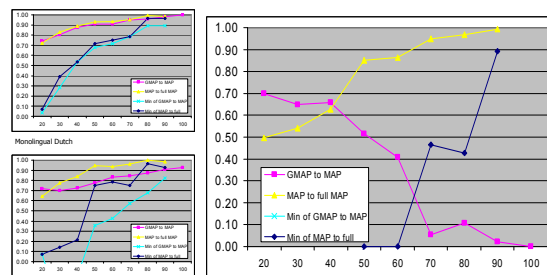
$$GMAP = \frac{1}{n} \sum_{i=1}^n \log x_i$$



CLEF Robust Task 2007



GMAP for multilingual runs



Conclusion GMAP


- GMAP measures something different than MAP
- These measures “measure somewhat different things” (Robertson 2006)
- However, it is not clear what they measure

Increasing Robustness

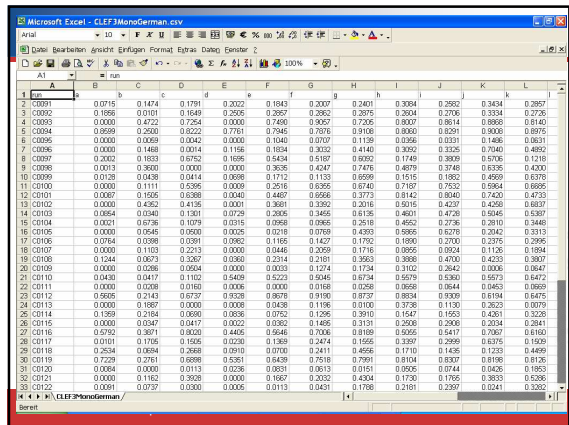
- Expand out of vocabulary terms using external collections (e.g. the Web)
- Word Sense Disambiguation

Activity

- Data Mining on evaluation results
 - explore real results from CLEF 2002
 - Runs per topic
 - Which aggregation leads to which ranking?



Mandl: Current Developments in Information Retrieval Evaluation




run	a	b	c	d	e	f	g	h	i	j	k	l
1	0.0001	0.0715	0.1474	0.1791	0.2002	0.1643	0.2007	0.2601	0.3054	0.2862	0.3434	0.2897
2	0.0002	0.1556	0.0701	0.1649	0.2092	0.2627	0.2862	0.2075	0.2604	0.2706	0.3334	0.2726
3	0.0003	0.0000	0.4722	0.7254	0.0000	0.7450	0.8057	0.7206	0.8007	0.8614	0.8986	0.8140
4	0.0004	0.9999	0.2000	0.0000	0.7351	0.7045	0.7076	0.6106	0.8050	0.8291	0.9006	0.8575
5	0.0005	0.0000	0.0009	0.8842	0.0000	0.1040	0.0707	0.1126	0.0366	0.0331	0.1486	0.0631
6	0.0006	0.0000	0.1468	0.0214	0.1156	0.1034	0.3002	0.4140	0.3262	0.3326	0.7040	0.4892
7	0.0007	0.2002	0.1853	0.6752	0.1696	0.5434	0.6167	0.6000	0.1749	0.3009	0.6706	0.1216
8	0.0008	0.0013	0.3600	0.0000	0.0000	0.3636	0.4247	0.7476	0.4679	0.3746	0.6336	0.4200
9	0.0009	0.0128	0.0436	0.0414	0.0698	0.1712	0.1132	0.6699	0.1516	0.1862	0.4669	0.6376
10	0.0010	0.0000	0.1111	0.6396	0.0009	0.2516	0.6395	0.6740	0.7197	0.7532	0.5984	0.6696
11	0.0011	0.0067	0.1056	0.6389	0.0040	0.4467	0.6666	0.3773	0.9142	0.9040	0.7420	0.4733
12	0.0012	0.0000	0.4562	0.4136	0.0001	0.3681	0.3392	0.2016	0.5016	0.4237	0.4256	0.6637
13	0.0013	0.0664	0.0340	0.1301	0.0729	0.2066	0.3466	0.6136	0.4611	0.4728	0.6046	0.5387
14	0.0014	0.0021	0.6736	0.1079	0.0315	0.0968	0.0966	0.2618	0.4552	0.2736	0.2810	0.3448
15	0.0015	0.0000	0.0646	0.0500	0.0025	0.0218	0.0769	0.4390	0.6966	0.6276	0.2342	0.3313
16	0.0016	0.0254	0.0396	0.0391	0.0962	0.1166	0.1427	0.1792	0.1900	0.2020	0.2376	0.2296
17	0.0017	0.0000	0.1103	0.2213	0.0000	0.0446	0.2059	0.1716	0.0966	0.0924	0.1126	0.1994
18	0.0018	0.1244	0.0673	0.5367	0.0040	0.2314	0.2191	0.3663	0.3668	0.4700	0.4230	0.3607
19	0.0019	0.0000	0.0286	0.0904	0.0000	0.0033	0.1274	0.1734	0.3102	0.2642	0.0006	0.0647
20	0.0020	0.0430	0.0417	0.1102	0.5409	0.5223	0.5045	0.6734	0.5579	0.5360	0.5573	0.6472
21	0.0021	0.0000	0.0008	0.0160	0.0006	0.0000	0.0168	0.0298	0.0658	0.0644	0.0463	0.0669
22	0.0022	0.6606	0.2143	0.6737	0.0000	0.8078	0.9100	0.8737	0.8834	0.9300	0.6194	0.6475
23	0.0023	0.0000	0.1867	0.0000	0.0006	0.0436	0.1196	0.0100	0.3736	0.1130	0.2623	0.0079
24	0.0024	0.1369	0.2184	0.0690	0.0636	0.0762	0.1296	0.3910	0.1547	0.1555	0.4261	0.3226
25	0.0025	0.0000	0.0047	0.0417	0.0022	0.0062	0.1495	0.3191	0.2638	0.2906	0.3334	0.2641
26	0.0026	0.5792	0.3871	0.8000	0.4405	0.8646	0.7006	0.9189	0.5055	0.5417	0.7067	0.6160
27	0.0027	0.0161	0.1765	0.0506	0.0230	0.1269	0.2414	0.1556	0.3287	0.2998	0.3575	0.1600
28	0.0028	0.2534	0.0654	0.2668	0.0910	0.0700	0.2411	0.4556	0.1710	0.1438	0.1223	0.4499
29	0.0029	0.7229	0.2761	0.6998	0.5261	0.6439	0.7618	0.7991	0.8104	0.8307	0.6198	0.6126
30	0.0030	0.0064	0.0000	0.0113	0.0236	0.0631	0.0515	0.0151	0.0505	0.0744	0.0426	0.1953
31	0.0031	0.0000	0.1162	0.3929	0.0000	0.1667	0.2032	0.4334	0.1730	0.1765	0.3833	0.5286
32	0.0032	0.0061	0.0737	0.0000	0.0005	0.0113	0.0431	0.1798	0.2191	0.2297	0.0241	0.3251

Evaluation initiatives

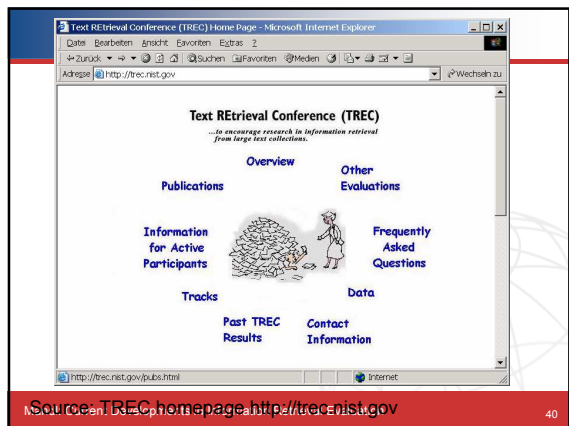
Overview

Aim:

- To establish a comparable basis for comparative evaluation



Mandl: Current Developments in Information Retrieval Evaluation



Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

Overview

- Publications
- Information for Active Participants
- Tracks
- Past TREC Results
- Other Evaluations
- Frequently Asked Questions
- Data
- Contact Information

Source: TREC homepage <http://trec.nist.gov>

Example topic

<TOP>
 <HEAD> Tipster Topic Description
 <NUM> Number: 066
 <DOM> Domain: Science and Technology
 <TITLE> Natural Language Processing
 <DESC>
 Document will identify a ctype of natural language processing technology which is being developed or marketed in the U.S.
 <NARR>
 A relevant document will identify a company or institution developing or marketing a natural language processing technology, identify the technology, and identify one or more features of the company's product.
 <CON> NLP, translation, language, dictionary, font, software
 <NAT> U.S.
 <TOP>

TREC-3: without CO
 TREC-4 and -5:
 only DESC
 from TREC-5:
 Varieties short and long

Quelle: Harman 1995: 8
 Mandl: Current Developments in Information Retrieval Evaluation

Example document: Wall Street Journal

<DOC>
 <DOCNO> WSJ880406-0090 </DOCNO>
 <HL> AT&T Unveils Services to Upgrade Phone Networks Under Global Plan </HL>
 <AUTHOR> Janet Guyon (WSJ Staff) </AUTHOR>
 <DATELINE> New York </DATELINE>
 <TEXT>
 American Telephon & Telegraph Co. Introduced the first of a new generation of phone services with broad ...
 </TEXT>
 </DOC>

Quelle: Harman 1995: 31
 Mandl: Current Developments in Information Retrieval Evaluation

Tracks in TREC

- Ad-hoc (until 2000)
- Filtering
- Spoken Language
- Question Answering
- Cross-Lingual -> CLEF in Europe
- Web
- ...

NIST

National Institute of Standards and Technology

Mandl: Current Developments in Information Retrieval Evaluation

43

Web-Track: Data

- Download of an extract of the internet
- Two corpora („web snapshots“)
 - small snapshot consists of 1,7 million pages (10 Gigabyte)
 - large snapshot consists of 18,5 million pages (100 Gigabyte)(Hawking 2001)
- Data are, unlike other tracks, with costs

NIST

National Institute of Standards and Technology

Mandl: Current Developments in Information Retrieval Evaluation

Web-Track: Method

- Quality problem
 - “assessors were asked to identify “best” documents for each topic” (Hawking 2001)
 - altogether less than 4% of the 70,000 found pages were classified as relevant
- Participating search engines “live” in the Web - Standard IRS in a controlled extract
 - more data for search engines (more relevant documents vs. more ballast)

Mandl: Current Developments in Information Retrieval Evaluation

Web-Track: Results

- In the first Web Track conventional Information Retrieval Systems performed better compared to tested internet search engines
- For the evaluation of search engines additional criteria has to be added (Crawler, degree of coverage, actuality, etc.)

NIST

National Institute of Standards and Technology

Mandl: Current Developments in Information Retrieval Evaluation

Web-Track: Results 2000

- PageRank
 - u.a. University of Neuchatel (know from CLEF) has tested PageRank in TREC web track
 - Hypertext-Links may improve the results of the Retrieval
 - That applies already to those in TREC used snapshots, that of course do not cover bei weitem the whole internet
 - **However improvement only for the retrieval of homepages**

Mandl: Current Developments in Information Retrieval Evaluation

Web-Track: Results 2001

- Significant differences of quality
- Homepage finding: 43 runs
 - Utilizing the linking structure and the internal structure of the documents improves the results
- Retrieval with regards to content: 77 runs
 - best system finds averagely 17 relevant documents among the first 100 hits

Mandl: Current Developments in Information Retrieval Evaluation

SPAM Track

- Spam Filter for e-mail and Retrieval

(Cormack 2006)

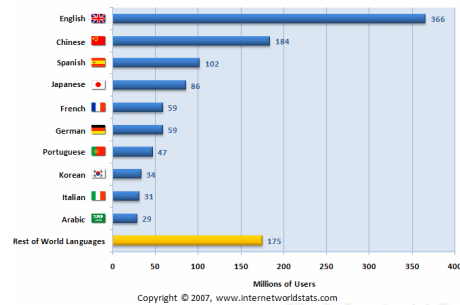
Tracks at TREC

- Enterprise Track
 - Search within heterogeneous Data of an organisation
- Genomics Track
 - Gene sequences and text documents
- HARD Track
 - High Accuracy Retrieval from documents
 - Additional information concerning seeker and using the context
- Question Answering Track
 - Answer concrete questions for facts
 - Shorter extracts (answers) from found documents extrahieren

Tracks

- Robust Retrieval Track
 - Focus on difficult queries
 - Not only the average of all topics counts
- Terabyte Track
 - Scaling up to mass data

10 Top Internet Languages



Cross-Language Evaluation Forum



For further information see:
<http://www.clef-campaign.org>

Charge:
Carol Peters - IEI-CNR, Pisa

Multilingual IR



- Query language differs from document language
- User situation:
 - User may understand documents in the foreign language (passive competence), but not being able to create a query (active competence)
- Access to documents
 - Dependent on relevance
 - Not dependent on language
- economic need
 - e.g. Patents in different languages

Multilingual Retrieval

- Query language differs from document language

Mandl: Current Developments in Information Retrieval Evaluation 55

Cross-Language Evaluation Forum

- Continuation of the Cross-Language Tracks by TREC for European languages
- Supply of an infrastructure for **encouraging** research and development in cross- and multilingual IR

Mandl: Current Developments in Information Retrieval Evaluation 56

Cross-Language Evaluation Forum

Challenges

- Multilingual corpus necessary (mainly from 1993 to 1994)
- Translation of queries in all available query languages
 - Maintain identical meaning
- native speaker for relevance judgment
 - Not the same person does all judgments for one topic

Mandl: Current Developments in Information Retrieval Evaluation 57

ORIGINAL LANGUAGE	TARGET LANGUAGE
EN "CNG cars"	DE "mit Flüssiggas betriebene Autos"
DE "Schneider-Konkurs"	FR „Faillite de M. Schneider"
NL "Muisarm"	FR "ordinateur: souris et tensions musculaire"
ES "Subasta de objetos de Lennon"	FR "Vente aux enchères de souvenir de John Lennon"
DE "deutsche Spätaussiedler"	EN "people of German origin from Eastern Europe coming to live in Germany"

Mandl: Current Developments in Information Retrieval Evaluation

Cross-Language Evaluation Forum

Newspapers and news agencies

- English: Los Angeles Times
- German: Schweizerische Depeschagentur-SDA, Der Spiegel, Frankfurter Rundschau
- French: Schweizerische Depeschagentur-SDA, Le Monde
- Italian: Schweizerische Depeschagentur-SDA, La Stampa
- Switzerland: NZZ
- Spain: Agencia EFE

Mandl: Current Developments in Information Retrieval Evaluation 59

Example for a topic

```
<top lang="EN">
<num>C086</num>
<EN-title>Renewable Power</EN-title>
...
<EN-narr>Relevant documents discuss the use of renewable energy sources such as solar, wind, biomass, hydro, and geothermal sources. Low emission vehicles as for example electric or CNG cars are not relevant. Fuel cells are not relevant unless their fuel qualifies as renewable.</EN-narr>
</top>

<top lang="EN">
num>C086</num>
<DE-title> Erneuerbare Energien </DE-title>
<DE-narr> Relevante Dokumente behandeln die Nutzung erneuerbarer Energiequellen, wie der Sonne, des Windes, der Biomasse, des Wassers und der Erdwärme. Schadstoffarme Fahrzeuge wie z.B. Elektroautos oder mit Flüssiggas betriebene Autos sind irrelevant. Brennstoffzellen sind nicht relevant, solange der Brennstoff nicht als erneuerbar gilt. </DE-narr>
</top>
```

Mandl: Current Developments in Information Retrieval Evaluation 60

Example for a topic

```
<top lang="ES">
<num>C083</num>
<ES-title> Subasta de objetos de Lennon. </ES-title>
<ES-desc> Encontrar subastas públicas de objetos de John Lennon. </ES-desc>
<ES-narr> Los documentos relevantes hablan de subastas que incluyen objetos
que pertenecieron a John Lennon, o que se atribuyen a John Lennon.</ES-narr>
</top>

<top>
<num>C083</num>
<FR-title> Vente aux enchères de souvenirs de John Lennon </FR-title>
<FR-desc> Trouvez les ventes aux enchères publiques des souvenirs de John
Lennon. </FR-desc>
<FR-narr> Des documents pertinents décriront les ventes aux enchères qui
incluent les objets qui ont appartenu à John Lennon ou qui ont été attribués à
John Lennon. </FR-narr>
</top>
```

Mandl: Current Developments in Information Retrieval Evaluation

61

```
<num> Number: AR13 <top> <num>C301</num>
<title> Théâtre en Égypte <PT-title>Marcas da Nestlé</PT-title>
<desc> Trouver les documents sur le théâtre en Égypte <PT-desc>Que marcas da Nestlé são
vendidas em todo o mundo?</PT-
desc>
<narr> Tout article concernant le théâtre égyptien, <PT-narr>Artigos relevantes devem
mencionar o nome de marcas
vendidas à escala global pela
Nestlé ou suas subsidiárias. No
segundo caso, devem fazer uma
referência clara à companhia-
mãe.</PT-narr> </top>
```

Mandl: Current Developments in Information Retrieval Evaluation

Tracks at CLEF

- **Ad-hoc**
 - Mono-lingual
 - **Persian, Czech, Hungarian, Bulgarian, English**
 - Bi-lingual
 - Amharic, Arabic, Oromo and Indonesian
 - Multi-lingual
 - **Portugese, German, English**

Languages used previously: Dutch, French, Spanish, Finnish, Swedish, Italian

Mandl: Current Developments in Information Retrieval Evaluation

63

Robust - Word Sense Disambiguation

Idea:

Provide English documents and topics (LA94 GH95) with automatically annotated word senses (WordNet)

Participants explore how the word senses (plus the semantic information in wordnets) can be used in (CL)IR

Tasks:

X2ENG and ENG2ENG

Mandl: Current Developments in Information Retrieval Evaluation

Example Document

```
<HEADLINE>
<TERM ID="GH950102-000000-1" LEMA="alien" POS="JJ">
<WF>Alien</WF>
<SYNSET SCORE="0.6" CODE="01295935-a"/>
<SYNSET SCORE="0.4" CODE="00984080-a"/>
</TERM>
<TERM ID="GH950102-000000-2" LEMA="treatment" POS="NN">
<WF>treatment</WF>
<SYNSET SCORE="0.827904118008605" CODE="00735486-n"/>
<SYNSET SCORE="0" CODE="03857483-n"/>
<SYNSET SCORE="0.172095881991395" CODE="00430183-n"/>
<SYNSET SCORE="0" CODE="05340429-n"/>
</TERM>
```

Mandl: Current Developments in Information Retrieval Evaluation

The European Library

Collection of catalog records

The TEL task used three collections:

- British Library (BL); 1,000,100 documents, 1.2 GB;
- Bibliothèque Nationale de France (BNF); 1,000,100 documents, 1.3 GB;
- Austrian National Library (ONB); 869,353 documents, 1.3 GB.

Main and expected language of the collection not the only language

multilingual, contains documents in many additional language

Mandl: Current Developments in Information Retrieval Evaluation

Tracks at CLEF

- **Domain Specific**
 - socio-scientific specialised texts
 - SOLIS Data base of the information centre Social Sciences
 - Russian corpus concerning Social Sciences
 - Thesaurus and intellectual indexing available
- **Image Retrieval**
 - Label of pictures („subline“)
 - Content Based: Medical pictures
 - e.g. radiographes
- **Spoken Document Retrieval**
 - Interviews
 - Challenge: Retrieval despite mistakes in speech recognition

GeoCLEF

- 20% of all Web queries have a geographic context
(Sanderson & Kohler 2004)
- **Geographical Information Retrieval (GIR)**
 - Queries with geographical reference
 - e.g. Articles concerning riots in the surrounding of Belfast
 - Definite geographical knowledge is combined with uncertain knowledge in the IR process

http://dx.doi.org/10.1007/978-3-540-85760-0_96

Example

```
<num>10.2452/89-GC</num>
<title>Trade fairs in Lower Saxony </title>
<desc>Documents reporting about industrial or cultural fairs in Lower Saxony. </desc>
<narr>Relevant documents should contain information about trade or industrial fairs which take place in the German federal state of Lower Saxony, i.e. name, type and place of the fair. The capital of Lower Saxony is Hanover. Other cities include Braunschweig, Osnabrück, Oldenburg and Göttingen. </narr>
```

Query GeoCLEF

```
<top>
<num>GC046</num>
<DE-title>Waldbrände in Nordportugal</DE-title>
<DE-desc>Dokumente über Waldbrände in Nordportugal</DE-desc>
<DE-narr>Dokumente sollen das Auftreten von, den Kampf gegen oder die Folgen von Waldbränden in Nordportugal berichten. Nordportugal umfasst die Regionen Minho, Douro Litoral, Trás-os-Montes und Alto Douro beziehungsweise die Distrikte Viana do Castelo, Braga, Porto (oder Oporto), Vila Real und Bragança.</DE-narr>
</top>
```

Queries

- „Forest fires in Northern Portugal“
- Where/what is Northern Portugal?
- Where is the boarder to Southern Portugal?
 - How high are the tolerances?
- Is there kind of Middle Portugal?



<http://www.wir-in-portugal.de/themen/landkarte.html>

Ambiguities

- Ambiguous: Galicia, Galicia (Spain, Poland)
 - Huge problem: Or (Stop word!)
- Different translations: Peking, Beijing
- Different names in different languages
 - Deutschland, Allemagne, Germany
- Change of name:
 - Bombay -> Mumbai
 - St. Petersburg -> Leningrad -> St. Petersburg

QA

- **Question Answering Track**
 - Answering concrete questions for facts
 - Extracting shorter abstracts (answers) from found documents
 - Eight languages as topic and document language possible

WebCLEF 2005-2006

- **Web Retrieval**
 - **First multilingual web-corpus**
 - Ca. 20 languages
 - 100 GB
 - **Exercises (short exercises with user model)**
 - Homepage Finding
 - Named Page Finding

WebCLEF

- For the first time 2005 EuroGOV corpus
- **European Regierungsseiten**
 - **Regierungsportal**
 - Websites of all ministries
- **27 domains sind vertreten**
- **25+ languages**
- **13 Main Domains with 131 ministries**

EuroGOV Collection Domains			
Main Domains	Additional Domains		
Domain	Country	Domain	Country
.cz	Czech Republic	.at	Austria
.de	Germany	.be	Belgium
.es	Spain	.cy	Cyprus
.eu	European Union	.dk	Denmark
.fi	Finland	.ee	Estonia
.fr	France	.gr	Greece
.hu	Hungary	.ie	Ireland
.it	Italy	.lt	Lithuania
.nl	The Netherlands	.lu	Luxembourg
.pt	Portugal	.lv	Latvia
.ru	Russia	.mt	Malta
.se	Sweden	.pl	Poland
.uk	United Kingdom	.si	Slovenia
		.sk	Slovakia

(De Rijke & Mishne, 2005:10)

Example: document

```
<EuroGOV:bin domain="se" id="001">
<EuroGOV:doc
url="http://www.regeringen.se/"
id="Ese-001-35"
md5="659b462005b40f04bde5946b2beaad71"
fetchDate="Wed Sep 22 10:57:39 MEST 2004"
contentType="text/html">
<EuroGOV:content>
<![CDATA[
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html lang="sv">
<head>
<title>Regeringen och Regeringskansliet</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<meta http-equiv="Content-Script-Type" content="text/javascript">
<meta http-equiv="Content-Style-Type" content="text/css">
<script language="javascript" type="text/javascript" src="/js/popup.js"></script>
<script language="javascript" type="text/javascript" src="/js/validationTexts_sv">
<script language="javascript" type="text/javascript" src="/js/formFunctions.js">
<link rel="stylesheet" type="text/css" href="/css/deprecatedstyle.css">

```

New: LogCLEF

Log Analysis and Geographic Query Identification (LAGI)

Similar to GeoCLEF 2007 subtask
Find geo queries in a web search engine log
Extract geo component

<http://www.uni-hildesheim.de/logclef/>

Task

Find queries with a geographic scope
Extract where component
Extract geo-relation-type
Extract what component
Classify what type {information, yellow page, map}

Example:

```
<local>YES
Lottery in Florida
<what>lottery
<what-type>information
<where>Florida, US
< geo-relation>in
```

Data 2007 and 2009

2007: Query log from the MSN search engine
mainly in English
800.000 queries (collected August 2006)
500 queries were labelled and used for evaluation
100 queries for training
400 for testing

2009: Query log from the TUMBA search engine
mainly in Portuguese
In cooperation with MITRE corp.

Mandl: Current Developments in Information Retrieval Evaluation

Negative Example

„Credit Card“ -> classified as Geo Query

Positive Training examples

„Credit River“
„Port Credit“
„Card Gulch“
„Card Bay“

Mandl: Current Developments in Information Retrieval Evaluation

80

New: LogCLEF

Log Analysis for Digital Societies (LADS)

TEL query and activity logs
analysis of user behavior

Can be done in combination with the TEL
content data

<http://www.uni-hildesheim.de/logclef/>

Mandl: Current Developments in Information Retrieval Evaluation

Log Analysis for Digital Societies

Query log analysis ...

to extract in an automatic way information on the use of facilities of the portal which are related to multilingual aspects and to geographic aspects
to see if it is possible to automatically extract "user digital communities" from the logs
in general, to see if it is possible to make clusters of users

Logs

The European Library (TEL) action logs from October 2007 till June 2008

Mandl: Current Developments in Information Retrieval Evaluation

81

Log Analysis for Digital Societies

Tasks

- Query reformulation
 - strategies for zero/too many result sets
- Multilingual search behaviour
 - use of different languages in the same session
- Personalization
 - clusters/patterns of users
- Geographical analysis
 - search for local and/or non local information

Leave participants to mine logs and give results/suggestions which can be used the following year
maybe giving some directions on what type of analysis

Mandl: Current Developments in Information Retrieval Evaluation

82

NTCIR



- Multilingual Retrieval for Asian languages
- Carried out in Tokyo
 - National Institute for Informatics
- Tasks
 - Cross-lingual
 - Chinese, Japanese, Korean -> English
 - Patent Retrieval
 - Web Retrieval
 - Question Answering

Mandl: Current Developments in Information Retrieval Evaluation

84

```

<DOC>
<DOCNO>CTS_ECO_0003302</DOCNO>
<LANG>CH</LANG>
• <HEADLINE>98年進出口皆大幅衰退 </HEADLINE>
<DATE>1999-01-08</DATE>
<TEXT>
<P>【記者謝錦芳台北報導】財政部七日公布去年全年海關進出口貿易統計，進出口皆出現少見的衰退幅度。出口比前年衰退九、四%，創下民國四十四年以來最嚴重的衰退；進口衰退八、五%，去年貿易出超五十九億美元，創下民國七十三年以來新紀錄，影響所及，去年經濟成長率恐怕無法達到五%。
<P>財政部統計長許國忠指出，受到亞洲金融風暴影響，去年我國對亞洲出口衰退幅度高達十八、五%，我國對日本貿易逆差則下一七六、九億美元的歷史新高。在亞洲景氣低迷之際，唯獨我國對歐洲出口仍有六、七%的成長率。若景氣在下半年逐漸復甦，今年出口成績可以由負轉正。
<P>根據財政部昨日公布資料顯示，去年十二月出口值為九十、六億美元，進口值九十三、五億美元，分別比前年同期衰退十三、八%、十一、八%，由於必需品與民船進口量突然增加十一億餘美元，出、進口相抵產生貿易逆差二、九億美元。統計長許國忠指出，去年十二月我國對香港、東協等地區出口減幅均超過廿五%，對美國、日本、歐洲的出口也呈現衰退現象。
<P>累計去年全年出口總值一、〇六、四億美元，進口總值一、〇四七、四億美元，雙雙突破一千億美元水準，進出口分別比前年衰退九、四%、八、五%，出口方面創下四十三年以來最嚴重的衰退。去年貿易出超萎縮至五十九億美元，也創下七十三年以來的新低。

```

Mand: <DOC>

Russian IR Evaluation


The Russian Information Retrieval Evaluation Seminar

- runs for the seventh time in 2009
- created evaluation resources for
 - newspapers
 - web documents
 - legal documents
 - collection of images from flickr.

<http://www.romip.ru/en>

Mand: Current Developments in Information Retrieval Evaluation

INEX



- **Initiative for the Evaluation of XML Retrieval**
 - Evaluation of retrieval with structured documents
 - Objective: not the whole document, but to find the smallest relevant extract
 - Using the structure
 - Test collection 2005: travel guide
 - <http://inex.is.informatik.uni-duisburg.de:2004/>

Mand: Current Developments in Information Retrieval Evaluation

FIRE

- Evaluation initiative for Indian languages
 - Hindi and Bengali (they belong to the 10 most spoken languages in the world)
 - Marathi, Tamil, Telugu, Punjabi, Malayalam
 - English
- Newspaper articles
- Started in 2008
- <http://www.isical.ac.in/~fire/>

Mand: Current Developments in Information Retrieval Evaluation

IMIRSEL

- **International Music Information Retrieval Systems Evaluation Laboratory Project**
 - Answering of W-questions concerning pieces of music
- <http://www.music-ir.org/evaluation/>

Mand: Current Developments in Information Retrieval Evaluation

TREC Video

Video Retrieval

<http://www-nlpir.nist.gov/projects/trecvid/>

Now also at:

Cross-Language Video Retrieval (VideoCLEF)

- Classification and retrieval tasks
- video collection of television program
 - in Dutch and English
 - Provided: speech recognition transcripts, metadata and shot-level keyframes
- Two tasks:
 - Subject Classification
 - Affect and Appeal

Mand: Current Developments in Information Retrieval Evaluation