Information Retrieval and Machine Learning

Massimo Melucci

University of Padua
Department of Information Engineering
massimo.melucci@unipd.it

CIMI School in Machine Learning 2015

- Introduction to Information Retrieval
  - Key Concepts
  - Nuts and bolts of indexing
  - Relevance Feedback
  - Evaluation
  - Towards modeling
- Information Retrieval Modeling
  - Boolean Modeling
  - Vector Space Modeling
  - Relevance Modeling
  - Language Modeling
  - Evaluation
- Machine Learning and Information Retrieval
  - Key Concepts
  - Correspondence between IR and ML
  - Feature (the input)
  - Approaches (the output)
  - Applications
  - Evaluation
- Hands-on Session
  - Data and Programs
  - Document indexing, ranking and evaluation

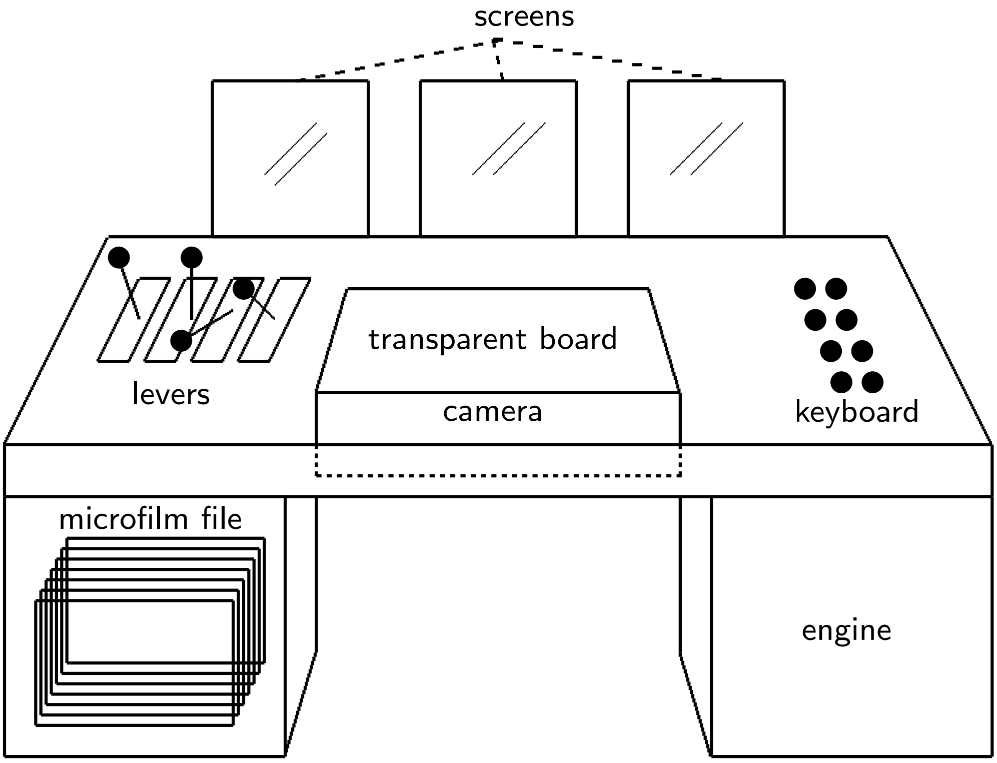
Memex

> *[S]cience may implement the ways in which man produces, stores, and consults the record of the race. [...] Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, memex will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory. [...]*
> *Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified.*

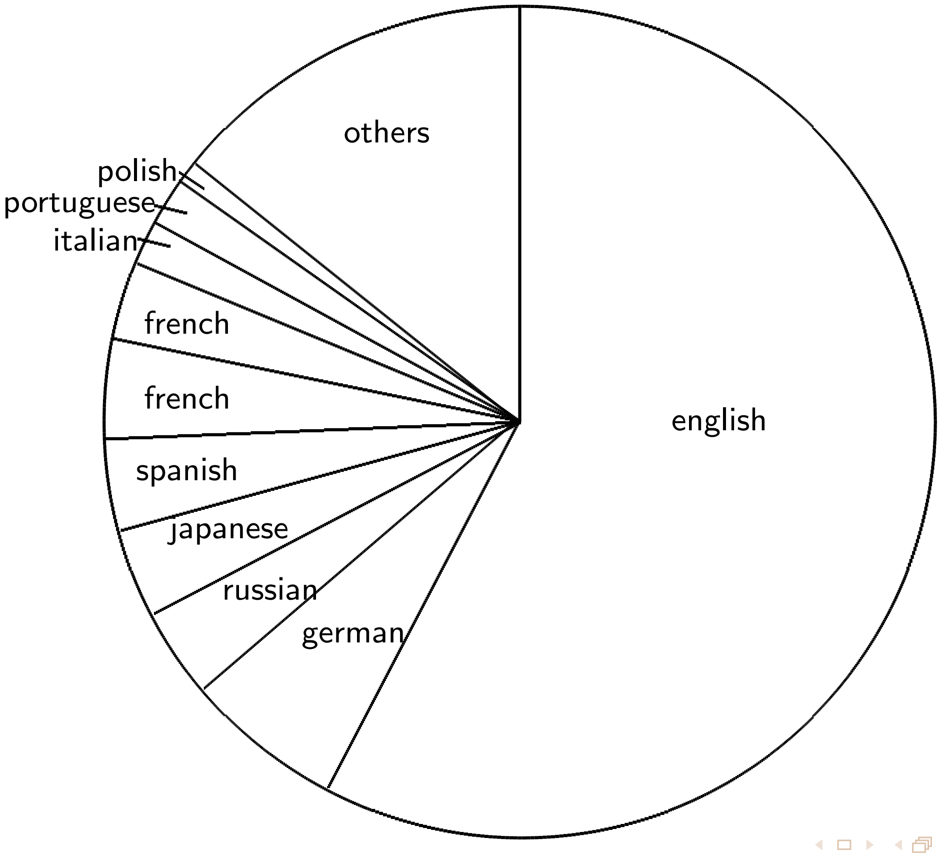
Bush

Bush V. (1945). As we may think.
Atlantic Monthly, **176**(1), 101–108.

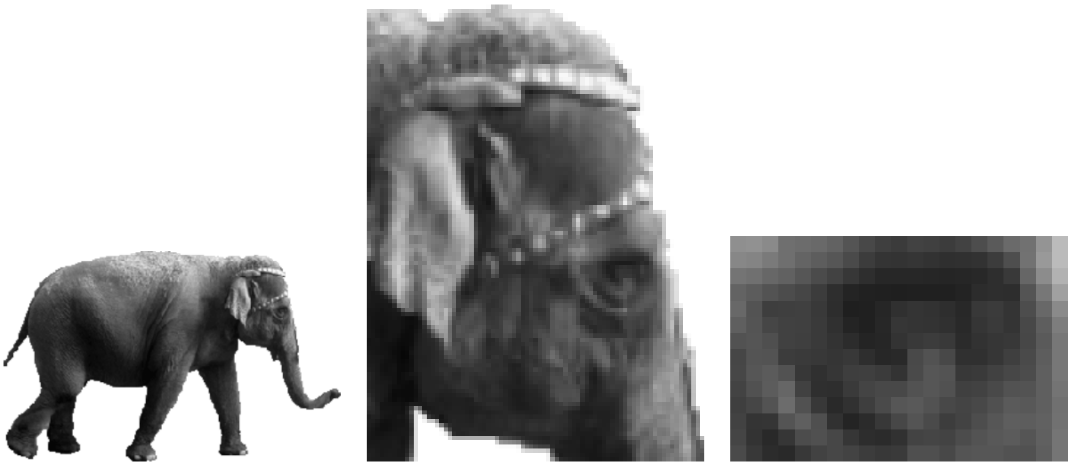
Memex in practice



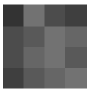
Languages



Images

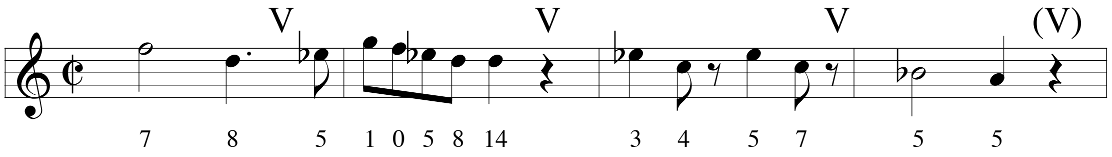


(a) full (b) head (c) eye



11	6	9	10
9	8	6	7
9	7	6	8
10	8	7	6

Music score



Information, data, need, and relevance

- Information: whatever changes user's knowledge to a degree necessary or sufficient that a task can be carried out or a problem can be solved.
- Data: symbols or signals that represent information.

  $$\text{Information} \neq \text{Data}$$

  - e.g. 90^o may refer to temperature, inclination, etc.
- Information need: information to solve user's problem or carry out user's task.
- Relevance: property of information that meet an information need.

  This is the *key* notion of Information Retrieval (IR).

Document, collection and query

- Document: container of data.
- Collection: container of documents.
- IR: activity that represents information as data and retrieve data that represents relevant information.
  - It might be automated yet not necessarily, in principle.
  - See also [2], [17], [44], [67], [75].
- Query: container of data that represent user's information need.
  - More later.

Ambiguity and languages

- Polysemy: property of words that have more meanings

$$\text{bank}\begin{cases}\text{Willows lined the bank of the stream.}\\ \text{Britain has a bank of highly exportable skills.}\\ \text{A bank of snow.}\end{cases}$$

- Synonymy: property of pairs of words that have the same meaning

$$\text{bank}\begin{cases}\text{Edge, side, embankment, levee, border...}\\ \text{Slope, rise, incline, gradient, ramp...}\\ \text{Financial institution, finance house, lender, mortgagee...}\end{cases}$$

- Multilingualism: socio-geo-ethnic-cultural case of synonymy
- Cross-language IR: IR with documents or queries / collections that contain data / documents in different languages

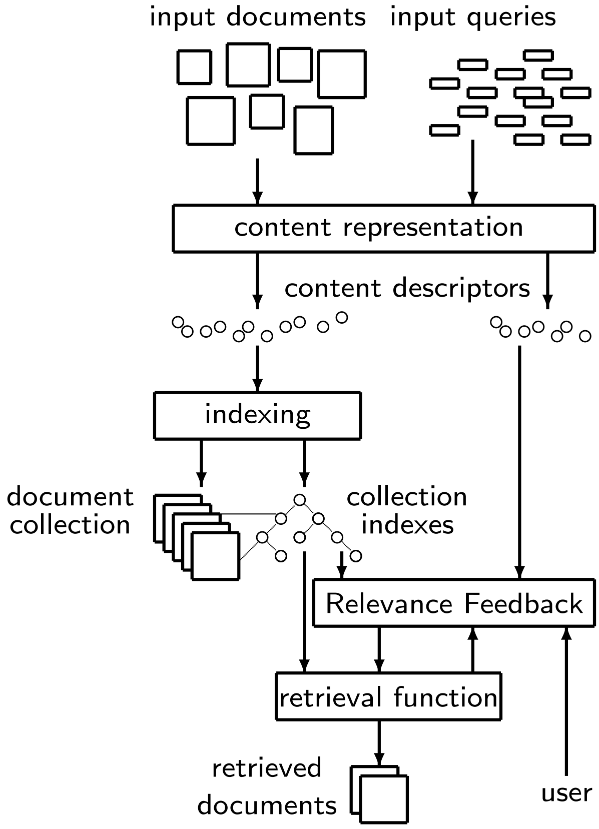
IRS

- IR System (IRS): computer system that performs IR
- Content descriptor: data about data (a.k.a. metadata)
- Term: textual content descriptor
- Posting: structure that relates term with documents
- Index: structure that stores postings
- Indexing: process that creates and updates indexes

Relevance Feedback (RF)

- Relevance Feedback (RF): process that updates information need descriptions using data observed during interaction
- Explicit RF.
- Pseudo Relevance Feedback (PRF).
- Implicit RF.

Information Retrieval System



Evaluation

- Evaluation: process that compares the effectiveness of two IRS that differ in e.g. two models
- Effectiveness: the degree to which an IRS is successful in producing relevant information.
- Efficiency: the quantity of the useful work performed by an IRS relative to the total computational cost (space, time, bandwidth) spent.

English stop-list

a
about
above
across
after
afterward
again
against
all
almost
alone
along
already
also
although
always
among
amongst

been
before
beforehan
behind
being
below
beside
besides
between
beyond
both
but
by
can
cannot
co
could
down

former
formerly
from
further
had
has
have
he
hence
her
here
hereafter
hereby
herein
hereupon
hers
herself
him

least
less
ltd
many
may
me
meanwhile
might
more
moreover
most
mostly
much
must
my
myself
namely
neither

only
onto
or
other
others
otherwise
our
ours
ourselves
out
over
own
per
perhaps
rather
same
seem
seemed

the
their
them
themselve
then
thence
there
thereafte
thereby
therefore
therein
thereupon
these
they
this
those
though
through

were
what
whatever
when
whence
whenever
where
whereafte
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who

English stop-list

an
and
another
any
anyhow
anyone
anything
anywhere
are
around
as
at
be
became
because
become
becomes
becoming
during
each
eg
either
else
elsewhere
enough
etc
even
ever
every
everyone
everythin
everywher
except
few
first
for
himself
his
how
however
i
ie
if
in
inc
indeed
into
is
it
its
itself
last
latter
latterly
never
neverthel
next
no
nobody
none
noone
nor
not
nothing
now
nowhere
of
off
often
on
once
one
seeming
seems
several
she
should
since
so
some
somehow
someone
something
sometime
sometimes
somewhere
still
such
than
that
throughou
thru
thus
to
together
too
toward
towards
under
until
up
upon
us
very
via
was
we
well
whoever
whole
whom
whose
why
will
with
within
without
would
yet
you
your
yours
yourself
yourselve

Stemming

- Stemming: word root identification.
  - computer, computing, computation: comput;
  - anti-fraud, defraud, fraudulence: fraud;
  - hide, hid, hidden: hid;
  - go, went, gone: ?
- Prefix: it is a special case of stem.
- Affix removal: affixes are listed in an affix table.
- Over-stemming: stem is longer than it should be.
- Under-stemming: stem is shorter than it should be.

Loss of information

	Testo A	Testo B
Original	In this paper we showed that the finiteness of the XSORT algorithm and that it always converge after a not large number of steps. We carried out experiments with large datasets.	This paper shows that the XSORT algorithm can not converge even after a large finite number of steps. Experiments have been carried out without large datasets.
After stop word removal	paper show finiteness xsort algorithm converge large number steps carried experiments large datasets	paper shows xsort algorithm converge large finite number steps experiments carried large dataset
After stemming	paper show finit xsort algorithm converg larg number step carri experiment larg dataset	paper show xsort algorithm converg larg finit number step experiment carri larg dataset
Bag-of-words	algorithm carri converg dataset experiment finit larg number paper show step xsort	algorithm carri converg dataset experiment finit larg number paper show step xsort

Lemmatization

Sistemi Informativi (corso progredito, Advanced Information Systems) is a graduate-level class in information retrieval offered at the University of Padua, Faculty of Statistics, Department of Information Engineering. This course covers the foundations of Information Retrieval and Search Engines as well as advanced or more recent topics. The lectures, homeworks, and laboratory assignments will in part be motivated by and organized around the design and implementation of a basic search algorithms useful in a real-world applications. This courses introduces the basics of Project Management for designing, implementing and evaluating search engine systems and algorithms. Core topics include material necessary to understand how an IR system is constructed and functions. The following topics will be covered: Indexing methods; Retrieval models; Web search engines; Machine learning; Evaluation. The course and the material will be in Italian. Further information: Massimo Melucci

Lemmatization

sistemi informativi (corso progredito / advance information system) be a graduate-level class in information retrieval offer at the university of padua / faculty of statistics / department of information engineer / this course cover the foundation of information retrieval and search engine as well as advance or more recent topic / the lecture / homework / and laboratory assignment will in part be motivate by and organize around the design and implementation of a basic search algorithm useful in a real-world application / this course introduce the basics of project management for design / implement and evaluate search engine system and algorithm / core topic include material necessary to understand how an ir system be construct and function / the follow topic will be cover / index method / retrieval model / web search engine ; machine learn / evaluation / the course and the material will be in italian / further information / massimo melucci

Part-of-Speech (POS) tagging

Sistemi Informativi (corso progredito, Advanced Information Systems) is a graduate-level class in information retrieval offered at the University of Padua, Faculty of Statistics, Department of Information Engineering. This course covers the foundations of Information Retrieval and Search Engines as well as advanced or more recent topics. The lectures, homeworks, and laboratory assignments will in part be motivated by and organized around the design and implementation of a basic search algorithms useful in a real-world applications. This courses introduces the basics of Project Management for designing, implementing and evaluating search engine systems and algorithms. Core topics include material necessary to understand how an IR system is constructed and functions. The following topics will be covered: Indexing methods; Retrieval models; Web search engines; Machine learning; Evaluation. The course and the material will be in Italian. Further information: Massimo Melucci

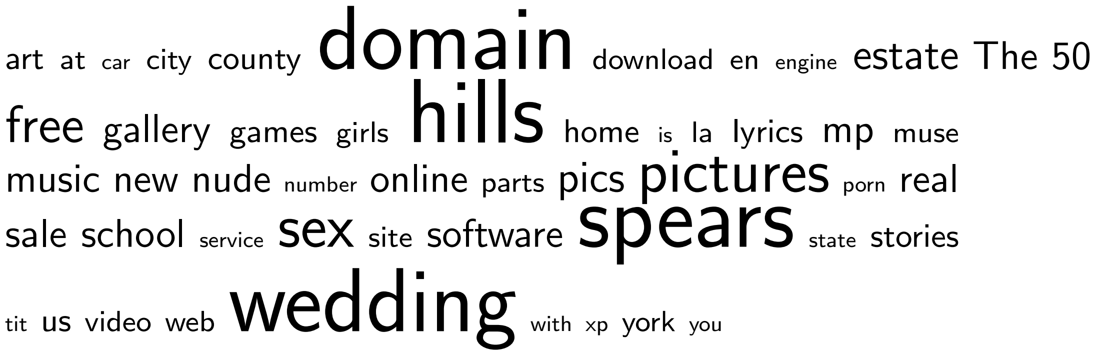
Part-of-Speech (POS) tagging

Sistemi/N Informativi/N corso/N progredito/N, Advanced/N Information/N Systems/N is/V a/I graduate-levelJ class/N in/P information/N retrieval/N offered/VD at/P the/D University/N of/P Padua/N, Faculty/N of/P Statistics/N, Department/N of/P Information/N Engineering/N. This/D course/N covers/V the/D foundations/N of/P Information/N Retrieval/N and/P Search/V Engines/N as/A well/A as/P advanced/V or/P more/A recentJ topics/N. The/D lectures/N, homeworks/N, and/P laboratory/N assignments/N will/V in/P part/N be/V motivated/V by/P and/P organized/VD around/P the/D design/N and/P implementation/N of/P a/I basicJ search/N algorithms/N usefulJ in/P a/I real-worldJ applications/N. This/D courses/N introduces/V the/D basics/N of/P Project/N Management/N for/P designing/V, implementing/V and/P evaluating/V search/N engine/N systems/N and/P algorithms/N. Core/N topics/N include/V material/N necessaryJ to/P understand/V how/A an/I IR/N system/N is/V constructed/V and/P functions/N. The/D following/V topics/N will/V be/V covered/V: Indexing/V methods/N; Retrieval/N models/N; Web/N search/N engines/N; Machine/N learning/N; Evaluation/N. The/D course/N and/P the/D material/N will/V be/V in/P Italian/N. Further/J information/N: Massimo/N Melucci/N

Again about queries and information need representation: queries *sensu lato* ranked by ambiguity

- SQL queries.
- Xpath queries.
- Regular expressions.
- Boolean queries.
- Bag-of-words.
- Clicks, dwell, save, eye movement.
- No action.

Distribution of query words

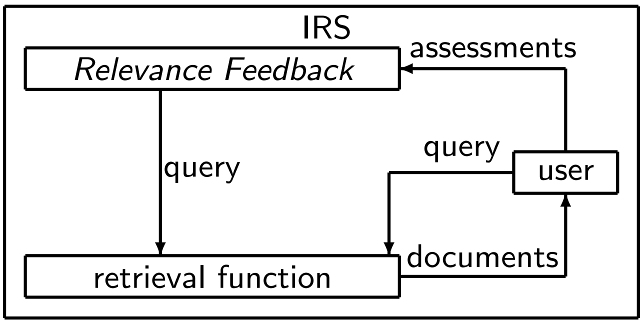


most frequent query words submitted to a search engine on 2002. [27]

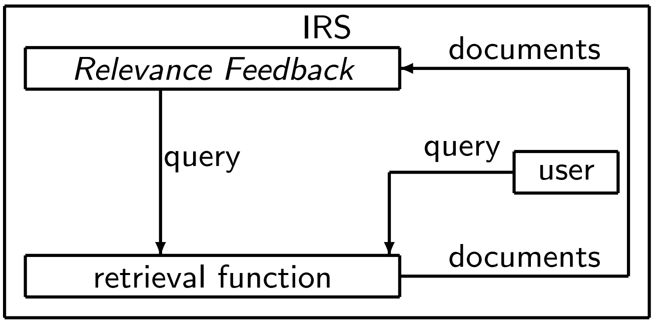
Summary

- Introduction to Information Retrieval
  - Key Concepts
  - Nuts and bolts of indexing
  - Relevance Feedback
  - Evaluation
  - Towards modeling
- Information Retrieval Modeling
  - Boolean Modeling
  - Vector Space Modeling
  - Relevance Modeling
  - Language Modeling
  - Evaluation
- Machine Learning and Information Retrieval
  - Key Concepts
  - Correspondence between IR and ML
  - Feature (the input)
  - Approaches (the output)
  - Applications
  - Evaluation
- Hands-on Session
  - Data and Programs

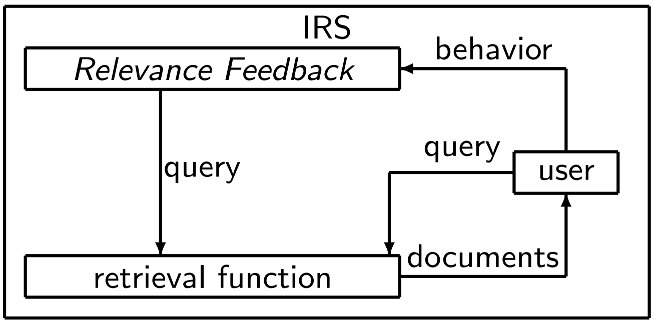
Explicit Relevance Feedback



Pseudo Relevance Feedback



Implicit Relevance Feedback



IRS

Relevance Feedback

behavior

query

query

user

documents

retrieval function

Summary

Introduction to Information Retrieval

- Key Concepts
- Nuts and bolts of indexing
- Relevance Feedback
- Evaluation
- Towards modeling

Information Retrieval Modeling

- Boolean Modeling
- Vector Space Modeling
- Relevance Modeling
- Language Modeling
- Evaluation

Machine Learning and Information Retrieval

- Key Concepts
- Correspondence between IR and ML
- Feature (the input)
- Approaches (the output)
- Applications
- Evaluation

Hands-on Session

- Data and Programs

What to evaluate

- IRS.
- IRS component.
- IRS configuration.
- Evaluation.
- Efficiency.
- Effectiveness.
- Relevance and context.

How to evaluate

- Study.
- Descriptive study.
- Explicative study.
- Explorative study,
- Laboratory study.
- Naturalistic study.
- User study.
- Longitudinal study.
- Case study.

How to evaluate

- Experiment: a scientific procedure undertaken to make a discovery, test a hypothesis, or demonstrate a known fact.
- Baseline: a minimum or starting point used for comparisons.
- Control group: an IRS configuration or user group that serves as a standard or reference for comparison with an experimental group. A control group is identical to the experimental group except that it is not subjected to experimentation.

Laboratory study

- Test collection.
- Document set.
- Topic set.
- Relevance assessment.

Measures

- Effectiveness measures.
- Efficiency measures.
- Precision.
- Recall.
- Gain.

MAP

- Average Precision (AP) of one ranking:

$$\frac{\mathrm{P}_1 + \cdots + \mathrm{P}_r}{r} \qquad \mathrm{P}_j = \frac{r_j}{n_j}$$

- Mean AP (MAP) of n rankings (e.g. queries):

$$\frac{\mathrm{AP}_1 + \cdots \mathrm{AP}_n}{n}$$

Mean AP

- For example, given two rankings:

	1	**2**	3	4	**5**	6	7	8	**9**	10
Precision	0	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{2}{5}$	$\frac{2}{6}$	$\frac{2}{7}$	$\frac{2}{8}$	$\frac{3}{9}$	$\frac{3}{10}$
Recall	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{3}{3}$	$\frac{3}{3}$

	1	**2**	3	4	**5**	**6**	**7**	8	9	10
Precision	1	$\frac{2}{2}$	$\frac{2}{3}$	$\frac{2}{4}$	$\frac{3}{5}$	$\frac{4}{6}$	$\frac{5}{7}$	$\frac{5}{8}$	$\frac{5}{9}$	$\frac{5}{10}$
Recall	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{5}{5}$

- We have

$$\mathrm{AP}_1 = \frac{1}{3}\left(\frac{1}{2}+\frac{2}{5}+\frac{1}{3}\right) = \frac{37}{90}$$

$$\mathrm{AP}_2 = \frac{1}{5}\left(1+1+\frac{3}{5}+\frac{2}{3}+\frac{5}{7}\right) = \frac{418}{525}$$

$$\mathrm{MAP} = \frac{1}{2}\left(\frac{37}{90}+\frac{418}{525}\right) = \frac{3803}{6300}$$

Cumulative gain

- gain from reading document at rank i: g_i
- cumulative gain from reading up to document at rank n

$$\mathrm{CG@}n = \sum_{i=1}^{n} g_i$$

- discounted gain from reading document at rank i

$$\mathrm{DG}_i = \frac{g_i}{i}$$

Cumulative gain

- discounted cumulative gain from reading up to document at rank n

$$\mathrm{DCG}@n = \sum_{i=1}^{n} \mathrm{DG}_i$$

- normalized discounted cumulative gain from reading up to document at rank n

$$\mathrm{NDCG}@n = \frac{\mathrm{DCG}@n}{\mathrm{DCG}^*@n}$$

where $\mathrm{DCG}^*@n = \max \mathrm{DCG}@n$

See [28].

Normalized Discounted Cumulative Gain (NDCG)

For example, given two rankings:

i	1	2	3	4	5	6	7	8	9	10
g_i	0	1	2	1	0	2	3	1	1	0
DG_i	0	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{4}$	0	$\frac{2}{6}$	$\frac{3}{7}$	$\frac{1}{8}$	$\frac{1}{9}$	0
DCG_i	0	$\frac{1}{2}$	$\frac{7}{6}$	$\frac{17}{12}$	$\frac{17}{12}$	$\frac{7}{4}$	$\frac{61}{28}$	$\frac{129}{56}$	$\frac{1217}{504}$	$\frac{1217}{504}$
Best case	3	2	2	1	1	1	1	0	0	0
DG^*_i	3	1	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$	0	0	0
DCG^*_i	3	4	$\frac{14}{3}$	$\frac{59}{12}$	$\frac{307}{60}$	$\frac{317}{60}$	$\frac{2279}{429}$	$\frac{2279}{429}$	$\frac{2279}{429}$	$\frac{2279}{429}$

We have that

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{DCG}^*} \approx 0.45$$

In one of most utilised versions, the discounted gain at rank i:

$$\mathrm{DG}_i = \frac{g_i}{\log_2 \mathrm{rank}_i + 1}$$

Evaluation campaign or initiative

- National Institute of Standard and Technology (NIST).
- Text REtrieval Conference (TREC).
- `http://trec.nist.gov`

Some test collections

Collection	Number of documents	Size	Average number of words / doc.
CACM	3,204	2.2MB	64
TIPSTER	500,523	6.4GB	740
.GOV2	25,205,179	426GB	1073
Clueweb09	1,040,809,705	25TB	32.2

TIPSTER Test collection

Source (Vol)	Year	Million words
Associated Press (1)	1989	40
Associated Press (2)	1988	37
Associated Press (3)	1990	37
Wall Street Journal	1987	20
	1988	17
	1989	6
Wall Street Journal (2)	1990	11
	1991	22
	1992	5
Dept. Of Energy (1)		28
Federal Register (1)	1989	38
Federal Register (2)	1988	30
Ziff/Davis (1)	1988	36
Ziff/Davis (2)	1989-90	26
Ziff/Davis (3)	1991-92	50
San Jose Mercury (3)	1991	45

A test document

Key documents prove innocence of Joseph Occhipinti – [in the U.S. District Court, Southern District of New York, criminal No. 91CR168 (CBM)]
(Extension of Remarks – November 04, 1993)
Hon. James A. Traficant, Jr., in the House of Representatives, Thursday, November 4, 1993
Mr. TRAFICANT. Mr. Speaker, for the last several months, I have been investigating the case of former Immigration and Naturalization Service Agent Joseph Occhipinti. Since his unjust... *to be continued*
The complete document is available at `http://thomas.loc.gov`

A test topic

<top>
<num> Number: 301
<title> International Organized Crime
<desc> Description: Identify organizations that participate in international criminal activity, the activity, and, if possible, collaborating organizations and the countries involved.
<narr> Narrative: A relevant document must as a minimum identify the organization and the type of illegal activity (e.g., Columbian cartel exporting cocaine). Vague references to international drug trade without identification of the organization(s) involved would not be relevant.
</top>

Run

topic	feedback iteration	document id.	rank	retrieval score	run tag
301	Q0	FBIS4-41991	1	-5.79809	indri
301	Q0	FBIS4-55395	2	-5.9537	indri
301	Q0	FBIS4-38364	3	-5.96086	indri
301	Q0	FBIS4-7811	4	-6.00289	indri
301	Q0	FBIS3-24143	5	-6.11308	indri
301	Q0	FBIS3-37418	6	-6.11528	indri
301	Q0	FBIS4-22471	7	-6.18484	indri
301	Q0	FBIS3-23986	8	-6.19521	indri
301	Q0	FBIS4-46734	9	-6.22884	indri
301	Q0	FBIS3-19646	10	-6.22924	indri

Relevance assessment

topic	feedback iteration	document id.	relevance assessment
301	0	CR93E-10279	0
301	0	CR93E-10505	0
301	0	CR93E-1282	1
301	0	CR93E-1850	0
301	0	CR93E-1860	0
301	0	CR93E-1952	0
301	0	CR93E-2191	0
301	0	CR93E-2473	0
301	0	CR93E-3103	1
301	0	CR93E-3284	0