

Оценка на търсещи машини

проф. д-р инж. Христо Вълчанов

<http://cs.tu-varna.bg>

Оценка на търсещи машини

Основен подход за изграждане на ефективни и ефикасни търсещи машини

- измерванията се извършват на базата на лабораторни експерименти;
- възможно е и он-лайн тестване.

Ефективност

- Измерване възможността на ТМ да намери правилната информация.

Рейтинговането, формирано от ТМ, да съответства на потребителските доказателства за съответствие.

Ефикасност

- Дефинира се в термините на изискванията за време и необходима памет за изпълнение на алгоритмите за рейтинговане.

Ефективност и Ефикасност

Изследванията акцентират основно върху подобряване на ефективността.

Вторичен е акцентът върху ефикасна реализация на тези техники.

Себестойност на ТМ

Изисквания <-> Себестойност.

Ограниченията за ефективност и ефикасност оказват влияние върху себестойността.

Ограниченията за ефикасност и себестойност могат да повлияят на ефективността.

Себестойност на ТМ

Пример за крайности при търсене в огромни колекции:

- *grep* – лоша ефективност и лоша ефикасност, но с ниска себестойност.
- библиотечни организации – високи ефективност и ефикасност (използване на специалисти), но много скъпо.

Оптимизация на ТМ

- Множество параметри, оказващи влияние.
- *Training data.*
- *Cost function.*

Тестови колекции

Тестовите колекции съдържат документи, запитвания, доказателства за съответствия.

- *CASM – заглавия и абстракти от Communications of ACM от 1958-1979. Запитванията и доказателствата са генерирани от компютърни учени.*

Тестови колекции

- *AP – документи на Associated Press от 1988-1990. Запитванията и доказателствата са генерирани от правителствени информационни аналитици.*

Тестови колекции

- *GOV2 – web страници, извлечени от сайтове в домейна .gov през 2004г. Запитванията и доказателствата са генерирани от правителствени информационни аналитици.*

Обем на колекциите

Collection	Number of documents	Size	Average number of words/doc.
CACM	3,204	2.2 MB	64
AP	242,918	0.7 GB	474
GOV2	25,205,179	426 GB	1073

Collection	Number of queries	Average number of words/query	Average number of relevant docs/query
CACM	64	13.0	16
AP	100	4.3	220
GOV2	150	3.1	180

Тестови колекции

- CACM – за търсещи приложения, фокусирани на библиографски записи. Доказателството е изчерпателно.
- AP – за пълен текст. Част от TREC.
- GOV2 – за web търсещи приложения. Част от TREC.

Примерно запитване при CASM

*Security considerations in local networks,
network operating systems and distributed
systems.*

TREC колекции

<https://trec.nist.gov/>



Application deadline to participate in TREC 2019 is now past.

[Celebration of the 25th TREC: November 15, 2016](#)

[TREC Economic Impact Study](#)

[TREC Statement on Product Testing and Advertising](#)

TREC колекции

<top>

<num> Number: 794

<title> pet therapy

<desc> Description:

How are pets or animals used in therapy for humans and what are the benefits?

<narr> Narrative:

Relevant documents must include details of how pet- or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it.

</top>

Журнали на запитванията

Използват се за настройка и оценка на ТМ.

Типично съдържание на журнал:

- ID на потребител или потребителски номер на сесия;
- Термини на запитването;
- Списък с резултата с URL, рейтинга им и кои от тях са кликнати;
- Времеви маркери – запомня се момента на изпращане на запитването, кликвания и др.

Информация за потребителското взаимодействие

- Използва се информация за кликванията (*clickthrough data*).
- *page dwell time* – времето, което потребителят е отделил на кликвания резултат;
- *search exit action* – начинът по който потребителят напуска приложението за търсене (затваряне на браузъра или изтичане на таймаут);
- разпечатване.

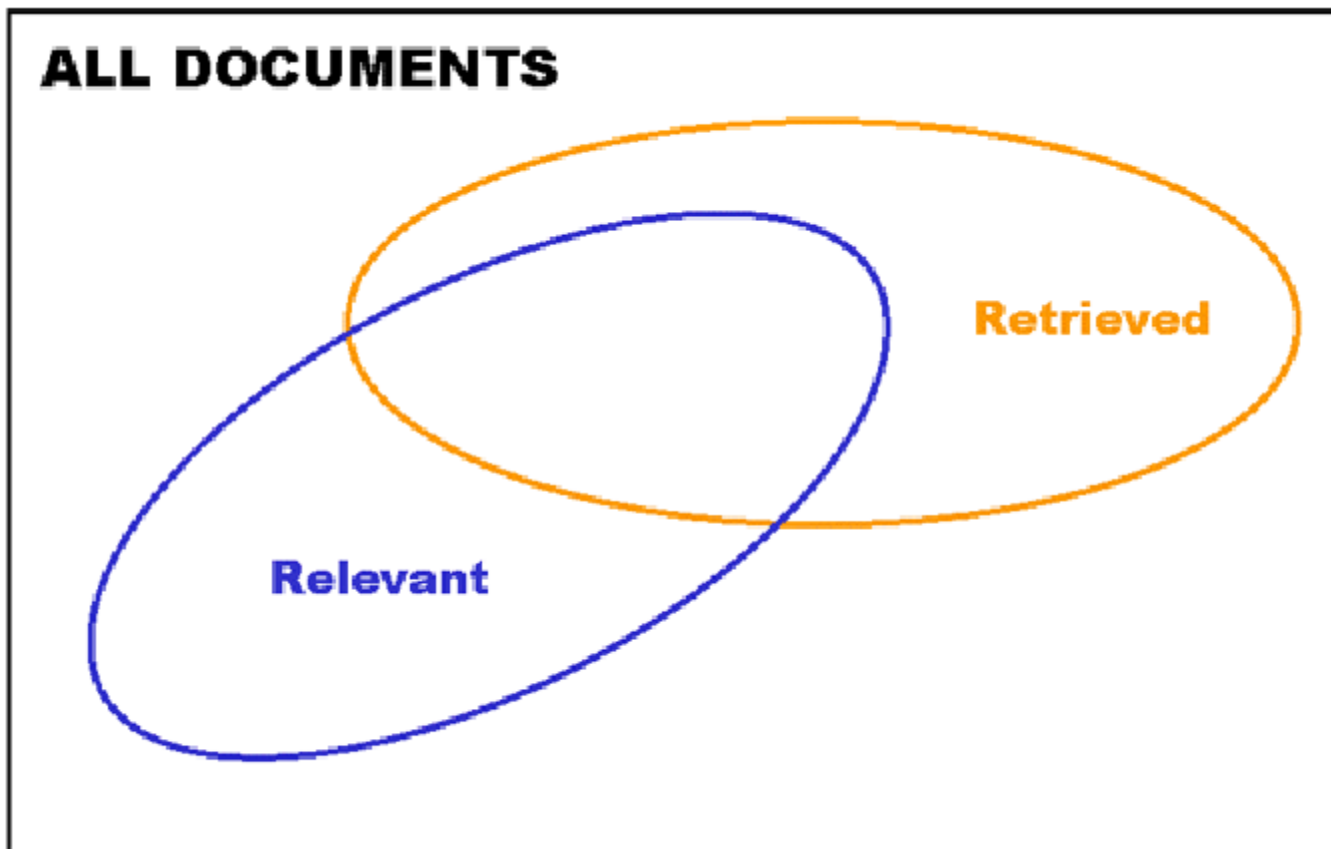
Метрики за ефективност

A – множество от релевантни документи

B – множество от извлечени документи

	Relevant	Non-Relevant
Retrieved	$A \cap B$	$\bar{A} \cap B$
Not Retrieved	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$

Метрики за ефективност



Метрики за ефективност

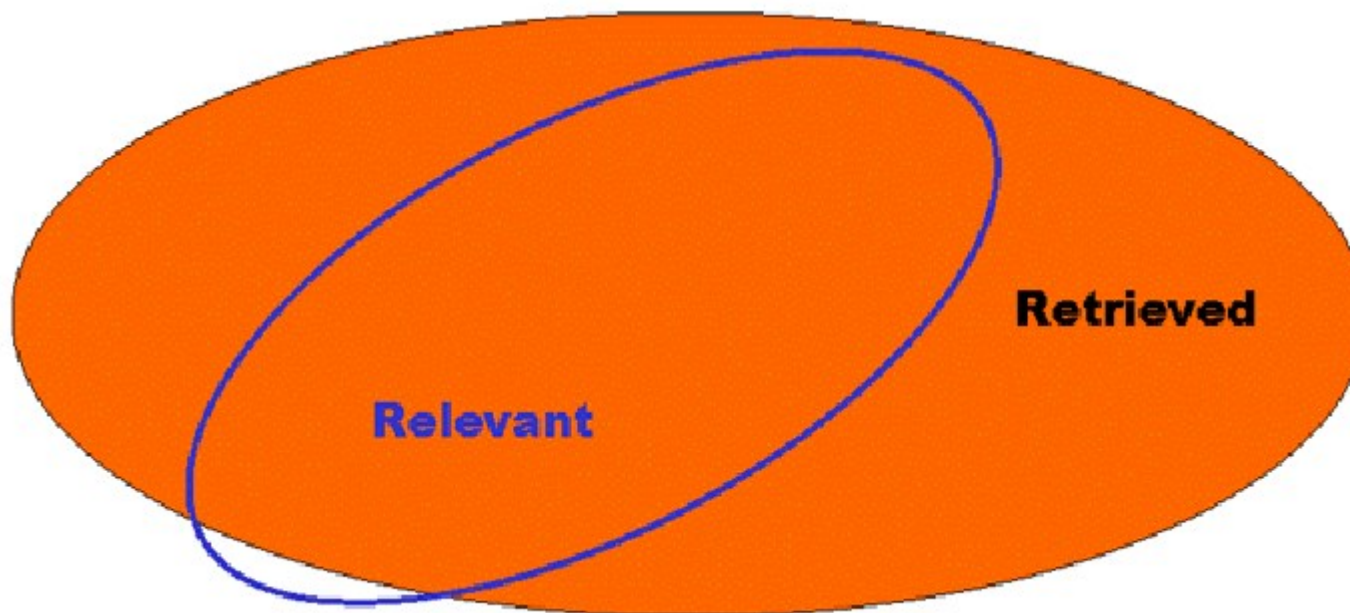
$$Recall = \frac{|A \cap B|}{|A|}$$

$$Precision = \frac{|A \cap B|}{|B|}$$

Recall е процента на релевантните документи, които са извлечени.

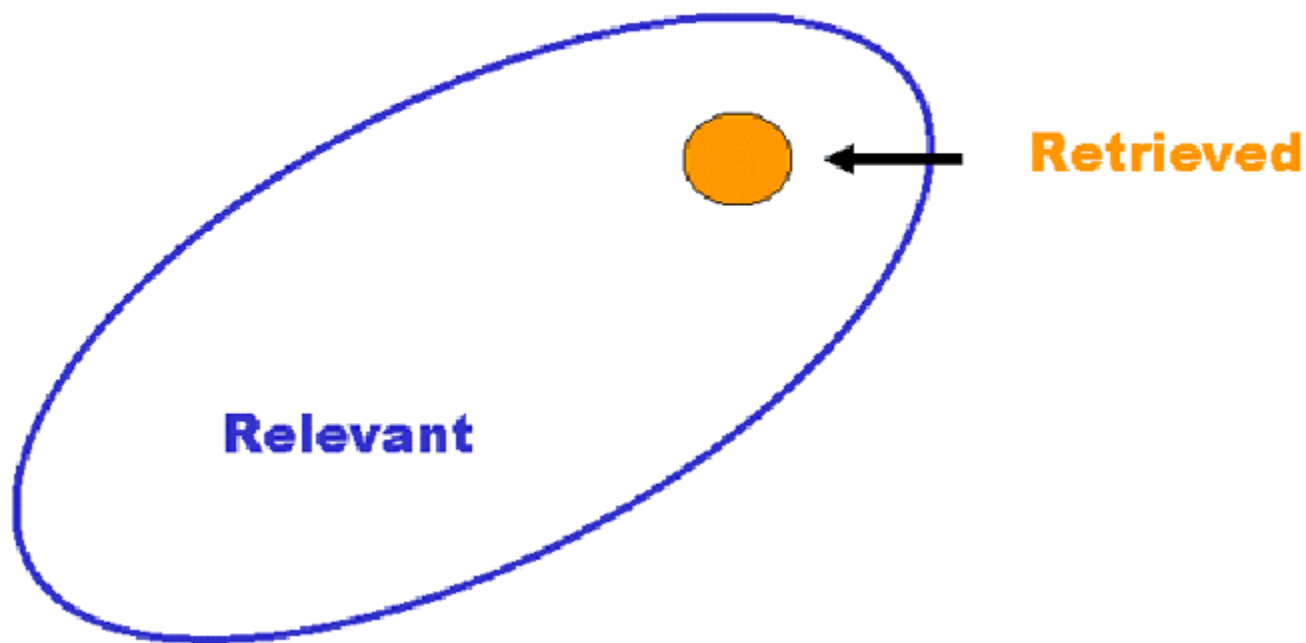
Precision е процента на извлечените документи, които са релевантни.

Съотношения



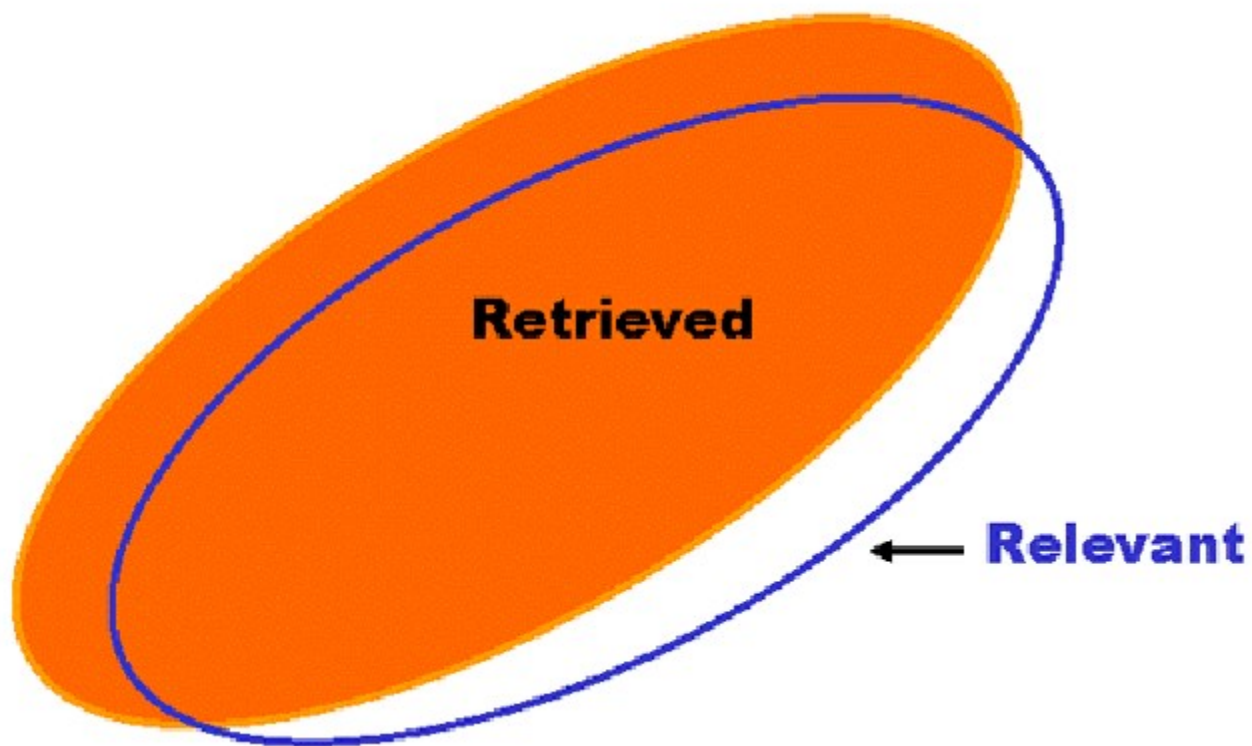
- *Recall* - висок
- *Precision* - нисък

Съотношения



- *Recall* - нисък
- *Precision* - висок

Съотношения



- *Recall* - висок
- *Precision* - висок

Метрики за ефикасност

- Време за индексиране;
- Процесорно време за индексиране;
- Пропускателна способност на запитванията;
- Латентност на запитванията;
- Временно пространство за индексиране;
- Размер на индекса.

Бърз отговор – до 150 ms

Латентност

99% от всички запитвания
да се обработят в рамките
на 100ms

Въпроси?