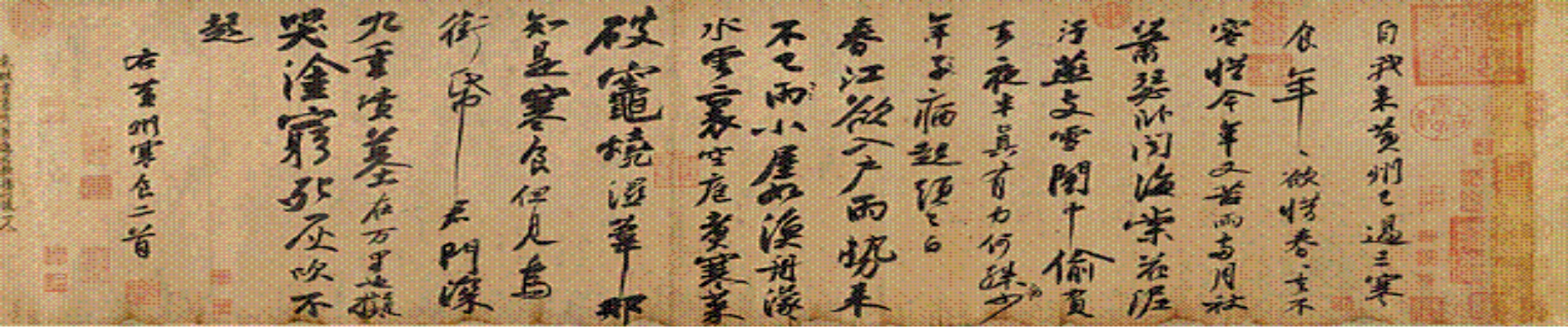


信息检索
Information Retrieval

教师：孙茂松

Tel:62781286，FIT 3-513

Email:sms@tsinghua.edu.cn

TA：**胡锦毅**

Email:hu-jy21@mails.tsinghua.edu.cn

郑重声明

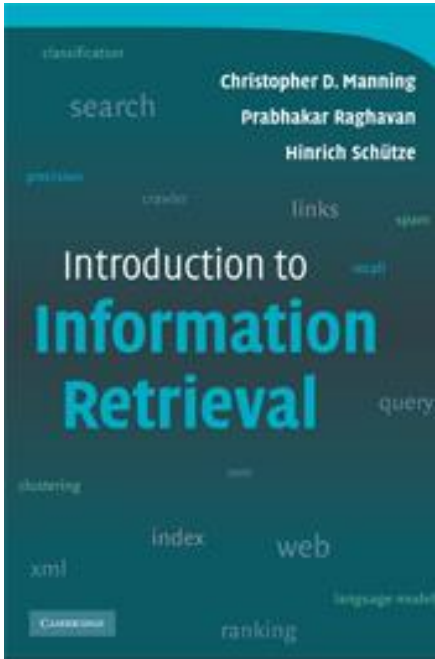
- 此课件仅供选修清华大学计算机系本科生课《信息检索》(40240372)的学生个人学习使用，所以只允许学生将其下载、存贮在自己的电脑中。未经孙茂松本人同意，任何人不得以任何方式扩散之。否则，由此可能引起的一切涉及知识产权的法律责任，概由该人负责。
- 此课件仅限孙茂松本人讲课使用。除孙茂松本人外，凡授课过程中，PTT文件显示此《郑重声明》之情形，即为侵权使用。

课程教材

- 教材：灵活

http://nlp.stanford.edu/IR-book/information-retrieval-book.html

1. Gerard Salton, Michael J. McGill, Introduction to modern information retrieval, McCraw-Hill International Book Company, 1983 (CALL NO: G354 ES17)
   1927-1995 co-founded the Department of Computer Science at Cornell University
2. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Introduction to Information Retrieval, Cambridge University Press, 2008.
3. 经典论文若干篇



参考资料

1. Text Retrieval Conference(TREC), http://trec.nist.gov/
   The TREC Conference series is co-sponsored by the NIST(an agency of the U.S. Commerce Department's Technology Administration), the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA) etc.
2. The International World Wide Web Conference, http://www.iw3c2.org/, http://www2002.org/, ...
3. Annual ACM Conference on Research and Development in Information Retrieval, http://portal.acm.org/

Text REtrieval Conference (TREC)

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

Overview

Frequently Asked Questions

Publications

Data

Information for Active Participants

Contact Information

Past TREC Results

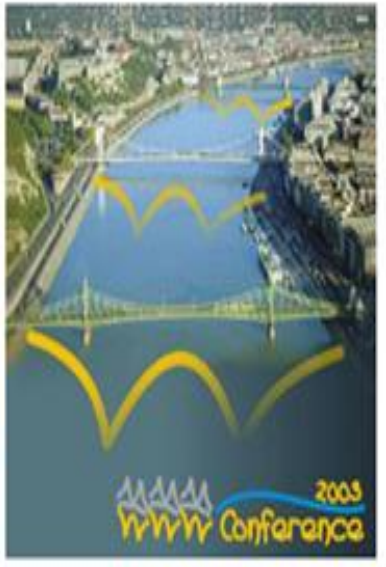
The TREC Conference series is co-sponsored by the NIST, Information Technology Laboratory's (ITL) Retrieval Group of the Information Access Division (IAD) and the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

http://WWW2003.org

The Twelfth International World Wide Web Conference

The Twelfth International World Wide Web Conference

Budapest Congress Centre
20-24 May 2003
Budapest, HUNGARY



Organisers

International World Wide Web Conference Committee (IW3C2)

Computer and Automation Research Institute of the Hungarian Academy of Sciences (MTA SZTAKI)

Search within SIGIR: Annual ACM Conference on Research and Development in Information Retrieval:

SEARCH Advanced Search

Browse SIGIR: Annual ACM Conference on Research and Development in Information Retrieval:

SIGIR

TOC Service

Current Year

2003 :Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval

Archive

1971 → 2002

The ACM SIGIR Conference focuses on research and development in information retrieval. It is the major international forum for the presentation of new research and the demonstration of new systems and techniques in the broad field of information retrieval.

进一步的读物……

Journal of the American Society of Information Sciences

ACM Transactions on Information Systems

Information Processing & Management

Information Retrieval

要求

- 成绩：作业(编程+DEMO) + seminar + 笔头作业 + 不定期的课堂小测验

  对IR不感兴趣的同学，请不要选。

  作业按时完成，否则该次作业记零分

- 一般不分组：(一人一组)

纪律：

- 鼓励讨论，严禁抄袭\拷贝：

  第一次发现，成绩记0分，并警告；

  第二次发现，整个课程不通过。

  （以上均无论抄与被抄者）。

- 不得迟到，早退，不得吃东西。关手机。

第一章 引言

1.1 从Vannevar Bush谈起

Vannevar Bush (March 11, 1890–June 30, 1974)

Internet Pioneers (& IR Pioneers)

http://graphics.cs.brown.edu/html/info/timeline.html

李开复:美国大学启示录 2004.9 http://tech.tom.com/1121/1793/2004922-127272.html

20世纪初美国的科研和大学仍然落后于欧洲。… 二战期间，在美国国家防务研究委员会主任Vannevar Bush的领导下，有六千名科学家机密地进行了大量的科研工作(包括影响深远的对原子弹、雷达、解密算法、导弹和青霉素的研究)。二战结束， Vannevar Bush调任国家科学研究与开发办公室主任。他提交给罗斯福总统一份名为《科学——无尽的战线》(“Science, the Endless Frontier”) 的报告，阐述他设计的一整套国家扶持科技（NSF，ARPA，ARPANET），利用科技创造财富的机制。

(On June 12, 1940, Bush met with President Roosevelt and detailed his plan for mobilizing military research. He proposed a new organization he called the National Defense Research Committee (NDRC). The committee would bring together government, military, business, and scientific leaders to coordinate military research. Roosevelt quickly agreed and thus the NDRC was created. Bush was made chairman and given a direct line to the White House).

Vannevar Bush不仅是政府官员，也是有独具慧眼的战略家和卓越的科学家。他在1931年研制成功的“微分分析仪”(Differential Analyzer)，是电子计算机的鼻祖。登月计划、SDI。他在1945年写的 “As We May Think”一文，预测了未来计算机、数据库、数位相机、语音识别、Internet等功能，有人因此称他为电脑之父。Vannevar Bush曾任麻省理工学院的副校长，曾创有名的Raytheon公司，帮助创立硅谷（ One of Bush's PhD students at MIT was Frederick Terman ），也是美国专利系统的创始人之一。

1.1 从Vannevar Bush谈起

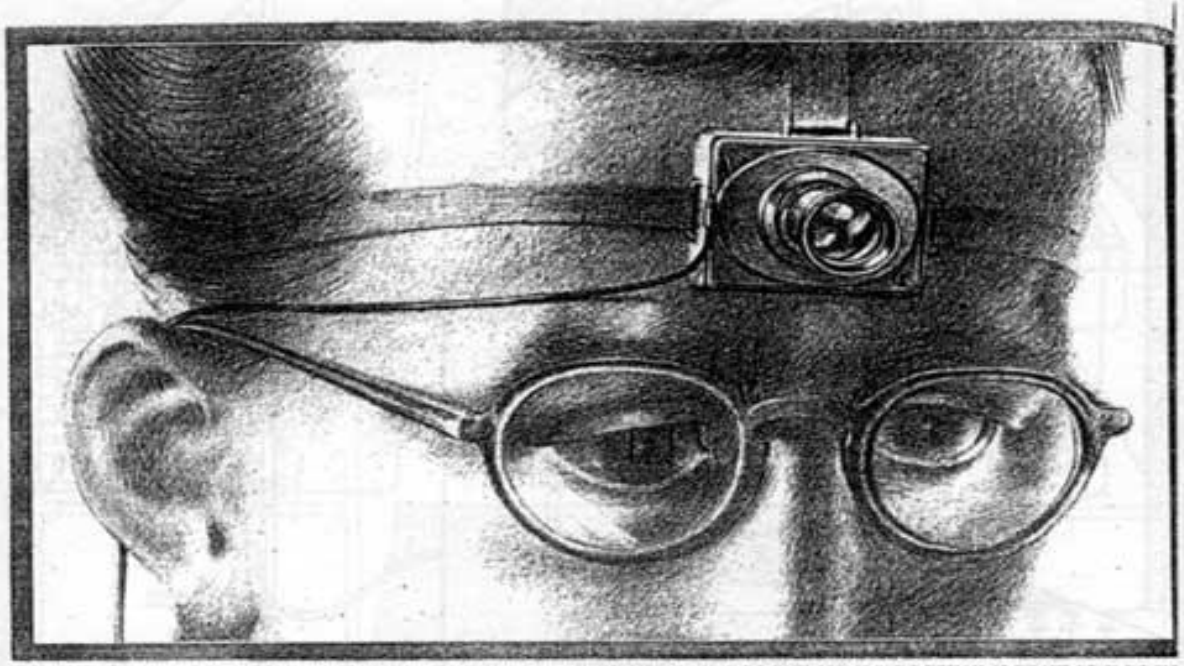
As We May Think by Vannevar Bush
Originally published in the July 1945 issue of The Atlantic Monthly.
http://web.mit.edu/STS.035/www/PDFs/think.pdf

For many years inventions have extended man's physical powers rather than the powers of his mind.

Now, says Dr. Bush, instruments are at hand which, if properly developed, will give man access to and command over the inherited knowledge of the ages.

This paper calls for a new relationship between thinking man and the sum of our knowledge.

Publication has been extended far beyond our present ability to make real use of the record.



A SCIENTIST OF THE FUTURE RECORDS EXPERIMENTS WITH A TINY CAMERA FITTED WITH UNIVERSAL-FOCUS LENS. THE SMALL SQUARE IN THE EYEGLASS AT THE LEFT SIGHTS THE OBJ

AS WE MAY THINK

A TOP U. S. SCIENTIST FORESEES A POSSIBLE FUTURE WORLD IN WHICH MAN-MADE MACHINES WILL START TO THINK

by VANNEVAR BUSH

DIRECTOR OF THE OFFICE OF SCIENTIFIC RESEARCH AND DEVELOPMENT

Condensed from the *Atlantic Monthly*, July 1945

This has not been a scientists' war; it has been a war in which all have had a part. The scientists, burying their old professional competition in the demand of a common cause, have shared greatly and learned much. It has been exhilarating to work in effective partnership. What are the scientists to do next?

For the biologists, and particularly for the medical scientists, there can be little indecision, for their war work has hardly required them to leave the old paths. Many indeed have been able to carry on their war research in their familiar peacetime laboratories. Their objectives remain much the same.

It is the physicists who have been thrown most violently off stride, who have left academic pursuits for the making of strange destructive gadgets, who have had to devise new methods for their unanticipated assignments. They have done their part on the devices that made it possible to turn back the enemy. They have worked in combined effort with the physicists of our allies. They have felt within themselves the stir of achievement. They have been part of a great team. Now one asks where they will find objectives worthy of their best.

* * *

There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers—conclusions which he cannot find time to grasp, much less to remember, as they appear. Yet specialization becomes increasingly necessary for progress, and the effort to bridge between disciplines is correspondingly superficial.

Professionally our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purpose. If the aggregate time spent in writing scholarly works and in reading them could be evaluated, the ratio between these amounts of time might well be startling. Those who conscientiously attempt to keep abreast of current thought, even in restricted fields, by close and continuous reading might well shy away from an examination calculated to show how much of the previous month's efforts could be produced on call.

Mendel's concept of the laws of genetics was lost to the world for a generation because his publication did not reach the few who were capable of grasping and extending it. This sort of catastrophe is undoubtedly being repeated all about us as truly significant attainments become lost in the mass of the inconsequential.

Publication has been extended far beyond our present ability to make use of the record. The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships.

But there are signs of a change as new and powerful instrumentalities come into use. Photocells capable of seeing things in a physical sense, advanced photography which can record what is seen or even what is not, thermionic tubes capable of controlling potent forces under the guidance

1.1 从Vannevar Bush谈起

The *Encyclopoedia Britannica* could be reduced to the volume of a matchbox. A library of a million volumes could be compressed into one end of a desk. If the human race has produced since the invention of movable type a total record, in the form of magazines, newspapers, books, tracts, advertising blurbs, correspondence, having a volume corresponding to a billion books, the whole affair, assembled and compressed, could be lugged off in a moving van. Mere compression, of course, is not enough; one needs not only to make and store a record but also to be able to consult it, and this aspect of the matter comes later.

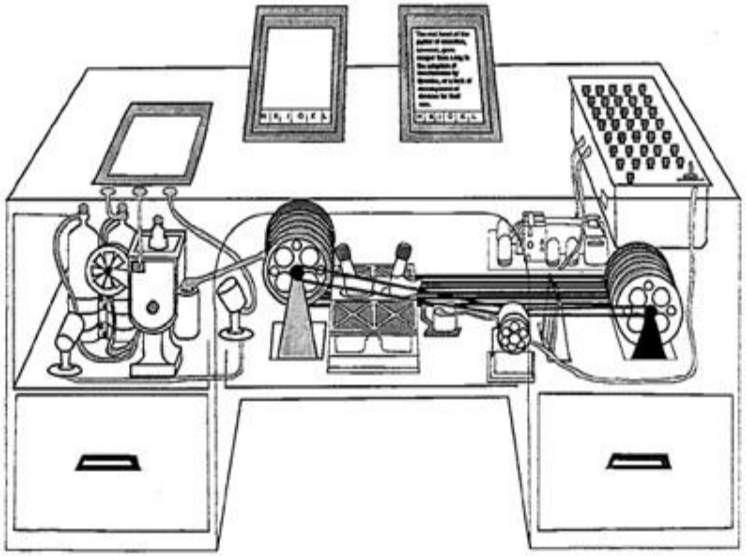
……

1.1 从Vannevar Bush谈起

Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage, they are filed alphabetically or numerically, and information is found (when it is) by tracing it down from subclass to subclass. It can be in only one place, unless duplicates are used; one has to have rules as to which path will locate it, and the rules are cumbersome. Having found one item, moreover, one has to emerge from the system and re-enter on a new path.

The human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts.

Memex



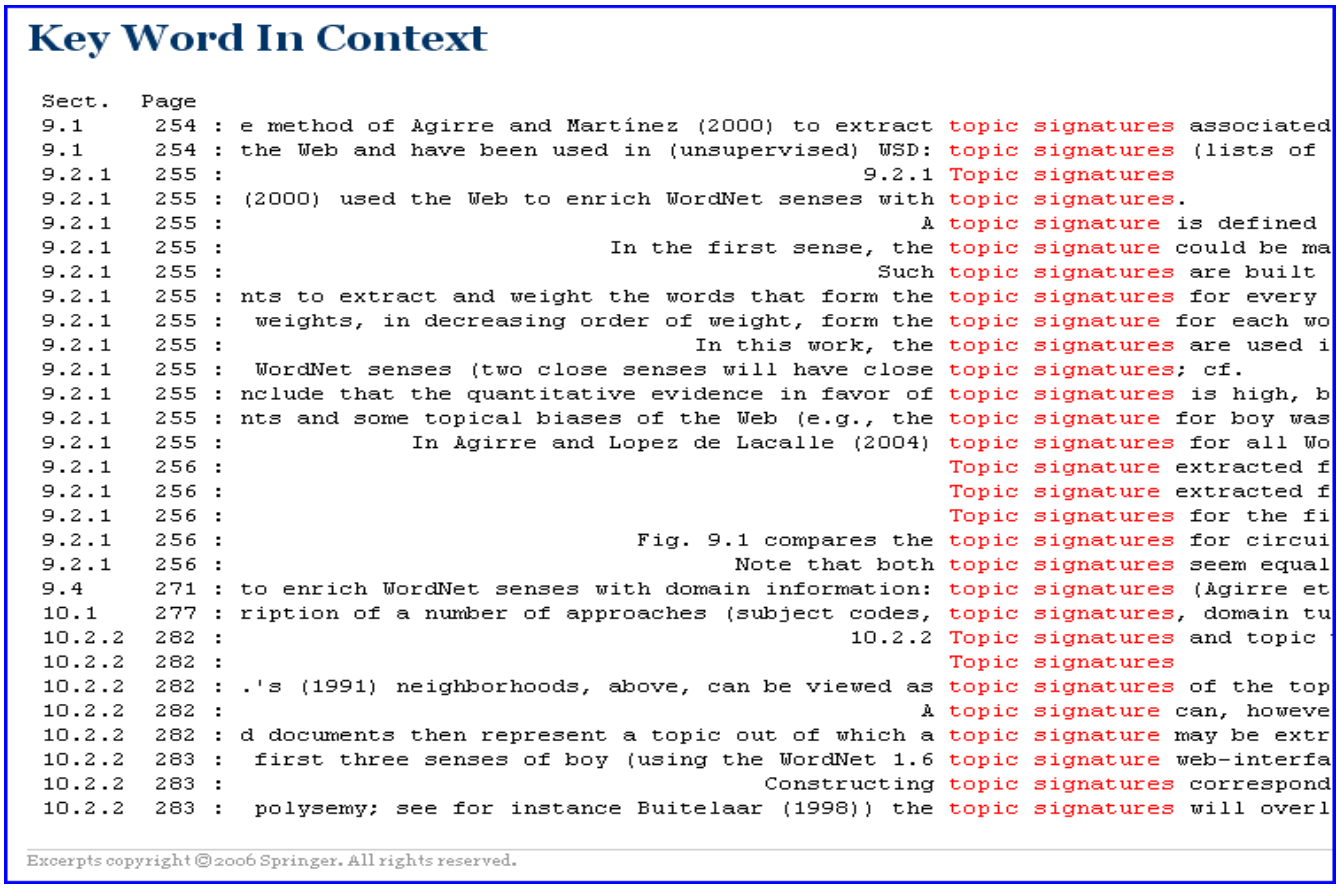
1.1 从Vannevar Bush谈起

This is similar to modern hypertext. In fact, Ted Nelson, who coined the term "hypertext" in the 1960's, acknowledges his debt to Bush. "Bush was right," says Nelson (Nelson in Nyce and Kahn, 245).

1.2 简史

（1）1945-1955

KWIC　Key Word In Context



Key Word In Context

Sect.	Page	
9.1	254	: e method of Agirre and Martínez (2000) to extract topic signatures associated
9.1	254	: the Web and have been used in (unsupervised) WSD: topic signatures (lists of
9.2.1	255	: 9.2.1 Topic signatures
9.2.1	255	: (2000) used the Web to enrich WordNet senses with topic signatures.
9.2.1	255	: A topic signature is defined
9.2.1	255	: In the first sense, the topic signature could be ma
9.2.1	255	: Such topic signatures are built
9.2.1	255	: nts to extract and weight the words that form the topic signatures for every
9.2.1	255	: weights, in decreasing order of weight, form the topic signature for each wo
9.2.1	255	: In this work, the topic signatures are used i
9.2.1	255	: WordNet senses (two close senses will have close topic signatures; cf.
9.2.1	255	: nclude that the quantitative evidence in favor of topic signatures is high, b
9.2.1	255	: nts and some topical biases of the Web (e.g., the topic signature for boy was
9.2.1	255	: In Agirre and Lopez de Lacalle (2004) topic signatures for all Wo
9.2.1	256	: Topic signature extracted f
9.2.1	256	: Topic signature extracted f
9.2.1	256	: Topic signatures for the fi
9.2.1	256	: Fig. 9.1 compares the topic signatures for circui
9.2.1	256	: Note that both topic signatures seem equal
9.4	271	: to enrich WordNet senses with domain information: topic signatures (Agirre et
10.1	277	: ription of a number of approaches (subject codes, topic signatures, domain tu
10.2.2	282	: 10.2.2 Topic signatures and topic
10.2.2	282	: Topic signatures
10.2.2	282	: .'s (1991) neighborhoods, above, can be viewed as topic signatures of the top
10.2.2	282	: A topic signature can, howeve
10.2.2	282	: d documents then represent a topic out of which a topic signature may be extr
10.2.2	283	: first three senses of boy (using the WordNet 1.6 topic signature web-interfa
10.2.2	283	: Constructing topic signatures correspond
10.2.2	283	: polysemy; see for instance Buitelaar (1998)) the topic signatures will overl

1.2 简史

（2）1960s

The late 1950s and 1960s were a time of great experimentation in information retrieval systems.

Early 1960s: Gerard Salton began work on IR at Harvard, later Cornell relevance feedback

Many of the commercial library systems such as Dialog can be traced back to experiments done at this time. The early 1960s also saw the definition of recall and precision and the development of the technology for evaluating retrieval systems.

Inspec Thesaurus

（3）1970s

Retrieval began to mature into real systems.

e.g., OCLC, the Online Computer Library Center

This decade saw the start of full-text retrieval systems

1.2 简史

（4）1980s

The widespread use of the CD-ROM.

Full text online blossomed in this decade.

（5）1990s

1989: First World Wide Web proposals by Tim Berners-Lee at CERN.

google PageRank

（6）2000s

Semantic Web, Knowledge graph

（7）2010s

Neural IR, OpenQA, VQA…

1.3 内涵

- 图书馆
  主题词表,手工编目,计算机检索,
  主题词自动标引,全文检索
- 数据库技术（结构化）
- WWW（自然语言文本+图象 Audio/Video,…）
  本课程主要针对（自然语言）文本信息检索（Text Information Retrieval) 非结构化
- 数字图书馆

1.3 内涵

- 文本信息检索主要研究方向

  文本信息检索 +搜索引擎

  （包括text, Audio,Image, Video检索等）

  文本自动分类、过滤

  （信息安全、网上不良品防护）

  自动文摘（包括关键词自动抽取）

  问答系统

  跨语言检索（机器翻译）

  文本挖掘（数据挖掘）

  Semantic Web

  知识管理

  海量数据处理（复杂性问题）

1.3 内涵

- 文本信息检索：跨学科
  计算机科学
  （网络支撑环境+
  自然语言处理+机器学习+深度学习 ）
  语言学
  统计学
  关键技术：自然语言处理（面向内容）