中文叙词表（用、代、属、分、参）

分词歧义

检全率、检准率、F值

自然语言文本

信息组织的原理

信息描述元数据作用、标准、MARC、XML、DDC

第四扎那边不用看

分类法结构、基本概述

各分类法的介绍不用看 -》细节안봐도됨

基本大类、部类、详表、复分表、简表

分类法原理、分类标引

主题法基本概念、类型、结构

主题法의 원리만 알면되고 자세한 표의 사용법은 안봐도됨

自然语言标引->讲义

교재 4，6,7,9,10안봄

四大模块 3,4집중적으로 복습

1-2

AS-PROCESS

知会的过程；对有关某些事实或现象的知识的交流；告知或被告知某些事实

AS-THING

传达信息的物件，例如数据或文档（文献），其 “具有灌输知识或交流信息的作用” 。

AS-KNOWLEDGE

过程中传播或交流的有关特定事实、主题或事件的 知识

组织的对象

信息资源——一切记录形式存在的信息载体 （p3)（information as thing)； 教材中常与“文献”(document) 通用

按基本类型，包括：书籍、报纸、论文、网 页、视频、音频、图片、 博客等各种类型；

按媒体的形式，包括：印刷型资源、电子资 源等。

组织的层次

1. 直接以信息资源存在单元为处理对象，如， 以图书、期刊、网站等为处理单元；

2. 以期刊、报刊、网站中的个体，如论文、新 闻、网站中的构成单元如网页等为处理单 元；

3. 直接以其中的信息成分作为处理的单元。

信息组织与信息检索

2-1

信息组织也是组织系统的一种

人和信息资源的交互有哪些

用户与信息资源的交互

访问和获取：最基本的交互，如借阅纸质图书、下载电⼦ ⽂档、在线观看电影视频等

查询/检索/问答：用户通过搜索引擎、数据库或图书馆目 录来查找特定信息：用户输⼊查询词，系统返回相关资源

浏览：非正式、⽆计划的多种信息搜寻活动

识别和选择：用户对信息资源进⾏辨识、判别和确认，从资源集合中选择信息资源

使用：阅读和观看、标记和复制，点赞、评论和分享等

广义信息系统的

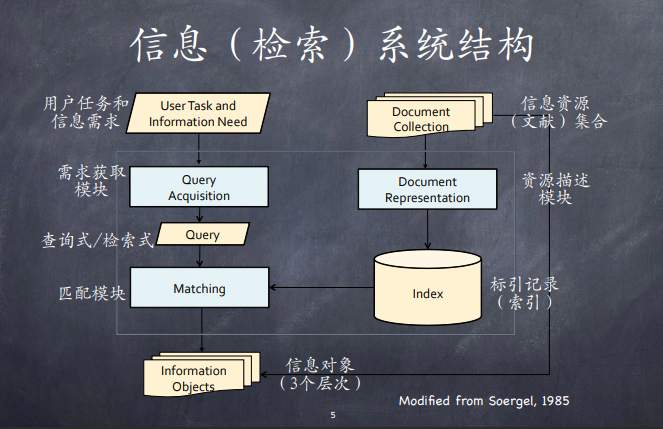
用户与信息资源的交互

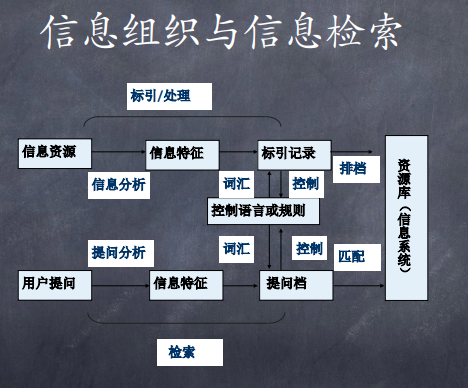
检索的类型

目录和索引은교재

2-2

信息检索系统结构





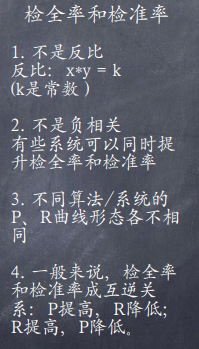
相关的概念

相关性 relevance

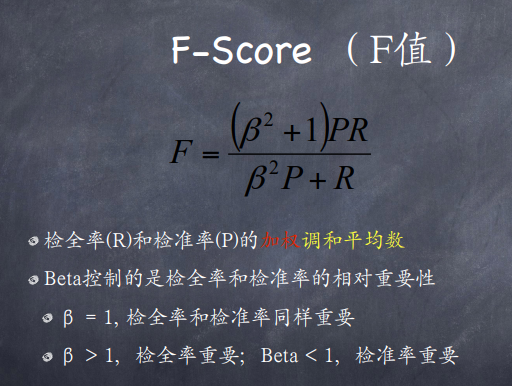
信息检索(Information Retrieval)：系统从信息资源（文献） 集合中，检索出与用户需求相关的信息对象的过程。

什么是相关性： 用户或系统对信息资源与用户需求之间匹配程度的一种判断。

检准率和检全率的计算



改良的计算



3-1

理论基础는 한번 훑어보기만 하면됨 별로 안중요한가봄

信息组织原则

文献保证原则——（知识资源）

用户保证原则——（用户）

与设备条件相适应——（中介）

文献保证

根据信息资源类型、数量等因素进行控制

资源类型：

图书的描述规范、网络视频资源的描述规范，描述的资源属性 有很大差别

图书分类、导航网站分类、视频网站分分类，在类目设置上大 相径庭

资源数量：

国家图书馆联机目录

家庭藏书目录

用户保证

根据用户需要进行描述控制和词汇控制

符合目标用户的基本特征（年龄、职业、受教育 程度等等）

符合用户的检索习惯（电商用户 vs 图书馆用户）

与设备条件相适应

手工 vs 计算机检索

显示屏幕大小

笔记本电脑 vs 平板 vs 手机

网络连接的速度

视频、图片的显示

与~相关的组织原则是什么

信息组织的活动도 별로 안중요

3-2

重点词汇控制

为什么需要进行词汇控制

词汇控制的目标

①单义性。即使概念与语词一一对应，一概念只用一语词表达， 一语词只表达一个概念。

②准确性。所使用语词应含义明确，具有科学、准确、通用的特点，不允许出现词义含糊现象。

③适用性。应适用于检索工具、符合用户使用习惯和文献状况。

④相关性。按使用需要，选择和规定词汇之间的联系，使语词之 间建立起严密的语义网络。

⑤兼容性

主题法랑같이보면됨

4-1

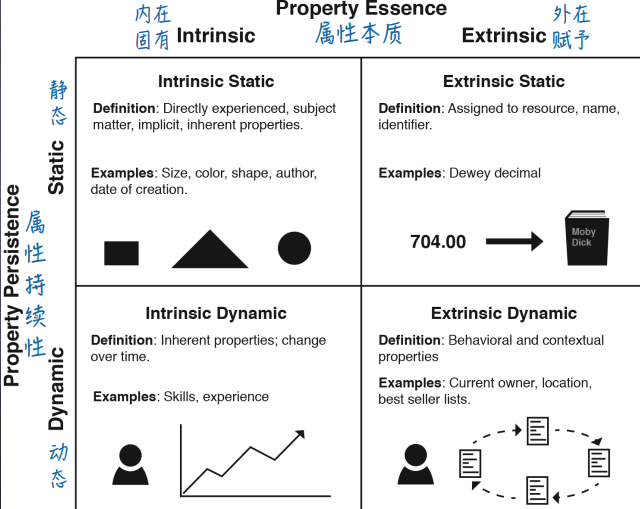
元数据与元数据的作用

元数据(meta data)——“data about data”

关于数据的数据，或关于数据的结构化的数据，是信息描述的结果

元数据通常是指从信息资源中抽取出来的用于说明其特征、内容的结构化的数据（如题名，版本、出版数据、 相关说明，包括检索点等）

用于组织、描述、检索、保存、管理信息和知识资源

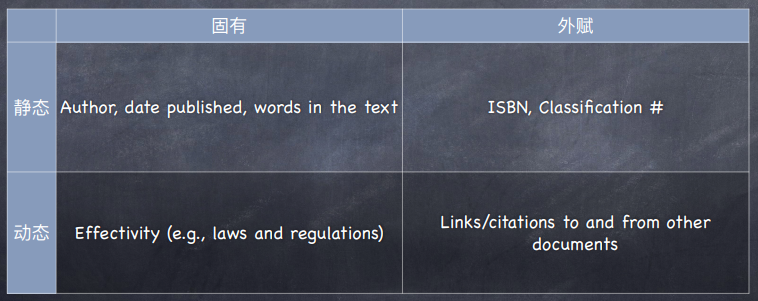


描述元数据——外部特征 Describes the information/data object and its properties

May use a variety of descriptive formats and rules Topical

主题元数据——主题内容 Describes the topic or “aboutness” of an information/ data object

May include a variety of vocabularies for describing, subjects, topics, categories, etc



著录项和检索点 《教材》

著录项/元素 entry

可以在目录或索引中找到的关于资源的描述项目或元素，例如作者名、题名、主题词等。

元数据规范要考虑的两个主要问题

哪些描述可以作为检索点? ——>检索点

著录项的内容是什么形式？（取值规则）—— 规范文

왜 4개의 표준이 있어야 하는지 이해

4-2

두 국가적 표준이 있는지 알아야함

《信息与文献资源描述》GB/T 3792-2021

都柏林核心集(Dublin Core) 【信息与文献 都柏林核心元数据元素集】GB/T 25100-2010

5-1

用户4个任务

Find entities that relate to a user's search criteria (locate and collocate)

Identify an entity, confirm it is what is sought, or distinguish it from similar

Select appropriate to users needs

Obtain the information entity through loan, or access electronically via a remote, networked computer

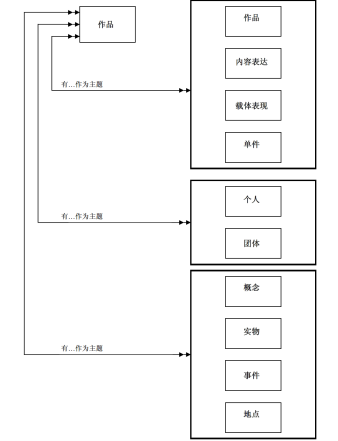
FRBR IN practice 보고 이해하기

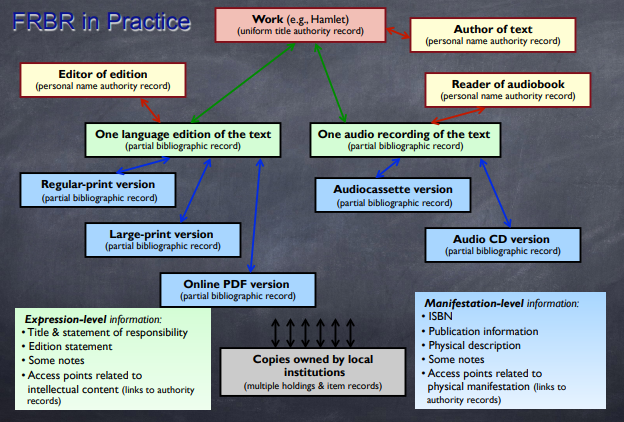
主要第一主

관계자세하게보기

作者与作品被创作关系

团体与作品被拥有的关系





5-2

Marc 자세히 안봐도됨

xml이랑 语义网이랑 같이 보면됨

如何实现互操作，为了系统间互操作

6-1

分类的基本的原理、基本的概念、困难

基本概念

类—指一组具

有某一共同属性的事物对象的集合。

分类—指根据事物的属性进行区分和类聚，并按照其相互关系进行组织的活动

包括两个基本方面：

其一，根据事物的属性区分或分组 ；

其二，是按照事物集合的关系确定类目的位置，并进一步按照其相同点和相异点区分和组织。

分类的困难

分类维度，属性相同，相异的考量

信息资源分类

信息资源分类——根据信息资源的内容属性和其他特征，将各种类型的信息资源分门别类地、系统地组织和揭示的方法。

信息资源分类特征：

是按照内容特征的相互关系加以组织的。

从一定的角度出发组织和揭示信息资源的。

以一定标记符号作为排序工具。

通过类目索引提供从字顺角度查找类目的途径

分类规则

惟一性： 在每一次划分时，只使用一个划分标准

完整性： 划分后所得各子类的外延之和应等于其母类的外延

科学性： 要选择事物本质的、符合分类目的的属性作为划分标准

类目的排列

同位类排列次序的重要性：

反映客观事物本身发展和联系的排列次序（自然序列）具有科学认识意义

符合某种具体目的的排列次序（人为序列）具有检索实用意义

类目的排列

对同位类排列的要求

连续性

规律性

稳定性

重要类目的突出性

连续性：所谓客观原则和历史原则，从简单到复杂、从低级到高级、照顾亲疏关系、照顾性质相近、位置相近、时间相近

类目名称及其涵义

类名涵义的理解

受上位类的限定

不仅受上位类的限定，而且也要受下位类的限定

除受上位类的限定外，还要借助于类目注释来明确

与其他有关类目进行对比

同一类目处于非最后一级和处于最后一级时的不同理解

类目名称要求

确切： 即所用的词或词组要能准确、恰当地反映类目的实际内容范围。

科学： 即要采用比较通行的科学名词作类目名称

简洁： 即所用的词或词组要概括精炼

解决的方案：

1.给定义

2.举个典型的例子

考虑题보기

6-2

等级列举和分面组配

等级列举

优势

类目体系直观，易于把握、便于使用

类目是经过配置的，展开比较系统

标记简明，适合分类排架和组织分类目录

不足

不能充分揭示细小专深主题

通常不能进行多角度检索

无法与科学的发展保持同步

篇幅较大

分面组配式分类法

一种按照分析－综合原则编制的分类表

分面组配式分类法只在列出各种基本概念（分面-facet）

使用时，先分析资源主题

根据主题分析的结果，通过相应分面的组配表达资源主题

特点

可通过概念组配，对复合主题进行专指标引

标记具有较强的表达性，便于调整组配次序，进行多元检索

可通过组配表达新出现的复杂主题，与科学的发展保持同步

类表的篇幅较小，便于控制、管理、增补、修订等

不足

类表不直观

类目的分布往往不均衡

标引难度较高

标记成分复杂，不适宜用于组织文献

列举－组配式分类法:일정한 등급제가 있고 다른 类랑도 조합이 가능한 방법

在详尽类表的基础上，广泛采用各种组配方式编制的分类法，亦称半分面分类法。

典型的列举组配式分类法－－ 《国际十进分类法》 (Universal Decimal Classification， 简称UDC)

大多数电商网站

7-1 교재랑 같이

中图法的例子需要了解

利用专类复分表进行复分

标记技术

预留空号法

即根据学科发展情况，在号码配置时，预先留下一些空号， 供类目增补时使用。如：

八分法

八分法——亦称扩九法，即在数字层累标记中，当同位类 超过10个，不足18个时，前９位以0－8表示，在9后面用两位数字表示一次划分，用于解决同位类的号码配置问题。

双位制

双位制——亦称集团标记法。即在数字层累制标记中，当同位类超过18个时，用两位数字表示一次划分，以解决号码的扩充问题。

借号法

借号法——即在层累标记制中使用的一种灵活借用上位类或下位类的号码配置方法。

通常在上一级号码较宽裕时，用上级号码标示下级类目，在下级类目超过１０个但超过不多时，根据情况，借用９以外的下级号码进行扩充或借用其它同级标记，使号码配置更加灵活。

字母标记法

8-1

分类体系的应用

分类导航被⼴泛用于：

展示资源或产品 (present resources/products)

支持对资源的浏览 (guide browsing)

控制对检索结果的展示 (structure search result display)

如何逻辑结构和展示怎么进行对应、映射

分类展示

以图书分类为例：分类排架——分类索书号

以⽹页分类为例：⽹页目录

以视频分类为例：视频分类导航

以商品分类为例：商品分类目录

9-2

主题法

主题法—直接以表达文献主题的语词作检索标识,以字顺为要检索途径的标引和检索信息资源的方法。 检索标识：内容主题词语

检索途径：字顺

主题法和分类法的区别



主题法的类型자세히

主题法的类型

按照是否控制

**自然语言主题法**

关键词法

特点： 可以直接使用计算机标引或处理，速度快，成本低，词汇更新及时；

专指性优于标题语言和叙词语言。

不足： 检索质量不稳定，容易受到文本表达质量的影响；

缺乏词汇控制。

中文关键词计算机抽词标引需要词表

**受控主题法**

词汇控制

词汇控制指根据标引和检索的需要，对自然语言的词汇进行选择、规范并揭示其相关性。

为什么要进行词汇控制？

词汇量过大——一些词标引无价值。

词汇与概念不一一对应——存在：

一义多词：计算机、电子计算机、电脑

一词多义：病毒－医学、计算机

词义含糊：软件分析－对软件进行分析、用软件分析？

缺乏明确的结构——自然语言词汇之间关系的多元性和不确定性，不符合检索 系统的使用特点

按照组配的先后

（词汇）组配

单个词语表示主题不够！标引缺乏精度。

需要通过组配来表示更专深的概念

如何解决复杂主题标引和检索的问题？

为先组式主题法:표기할때 어떠한 방식으로 조합을해서 표기를 하는방식

先组式检索语言——指在标引阶段，复杂主题的标识在系统中就已经按一定方式组配好了的一种标引和检索的语言。包括：标题法；叙词法也可编制先组式检索工具。

后组式主题法 ：따로분해한후 검색할때 검색언어로 조합

后组式检索语言——指检索前，复杂主题的标识在检索系统中是以基本概念的形式独立存在的，实施检索时，才根据检索提问进行组合检索语言。包括：元词法、叙词法。

按照选词方式

标题法

标题法 —一种以标题词作为主题标识，以词表预先确定的组配方式标引和检索的主题法。

标题词—亦称标题，指经过规范化处理的，用来标引文献的词或词组，通常为比较定型的事物名称。

标题法的特点与不足

特点

采用列举式词表,形式直观

定组式标题结构固定，含义明确

按照词表列举的标题和副标题进行标引，操作简便

主要通过以参照方式揭示标题之间的相关性

不足

收词量大

专指度不足

修订量大

定组式标题， 手工系统中不易多元检索

元词法

元词法—以元词作为主题标识，通过字面组配的方式表达文献主题的主题法。

元词—最基本的、字面上不能再分的语词。

例： 文献分类、主题标引——非元词； 文献、分类、主题、标引——元词。

字面组配，例：

“经济文献检索”，标引为：经济、文献、检索 “生物物理学”，标引为：生物、物理学

叙词法

叙词法——以叙词作为文献主题标识，通过概念组配方式表达文献主题的主题法类型。

叙词——亦称主题词，是经过规范化处理的，以基本概念为基础的表达文献主题的词和词组。

**特点**

概念组配，能准确揭示主题。

通过参照系统和索引，对词汇有效控制。

多途径检索，可采用灵活组配方式及对词间联系充分揭示检索。

同时适用于多种方式，包括：标识单元和文献单元检索方式，机检和手检

**不足**

词表编制和管理难度大，花费较多人力、物力；

标引在概念分析的基础上进行，标引难度大 。

10-1

主题检索 -重要

索引——顺排档



按资源的信息进行描述，组织的

不适合进行检索，适合于查看资源，

索引——倒排档

适合用于检索，用主题的指向资源

主题索引、

主题检索功能 - >借助受控词表进行的

**主题词检索功能**。通过输入主题词，输出命中的检索对象，命中篇数，及相 关信息。

**入口词检索功能**。输入与主题词具有用代关系的词，按对应主题词输出检索 结果。

第一步：找到入口词对应的叙词

第二步：用叙词进行检索

**组配检索功能**。结合逻辑组配符，进行组配检索，扩大或缩小检索范围，改 进检索效果。

**查询词扩展功能**。将用户的查询词进行上下位词、相关词扩展，提 升检索效果。

**综合查找功能**。可结合其他描述数据进行综合查找。

例，北大图书馆联机目录的“复杂检索”可同时结合主题、著者、 题名等进行检索。

**浏览查找功能**。利用预先确定的主题词（如标题词、叙词等）进行 浏览查找。

人工标引的结果怎么匹配上

10-2

词汇选择、词型控制제외한 2개 보면됨

词义控制

同义控制：找出同义/准同义 词串，选出优选 词/叙词

多义控制：找出可能存在歧义的词，加以限定

词间关系（用代属分参）每一类的作用、什么时候使用

等同关系

等同关系—亦称用代关系，指叙词与非叙词之间的关系。

作用—集中同一主题文献、增加检索入口、控制专指度。

使用范围—在同义规范基础上进行，应有参照和反参照

送别 （优选词，더 정식적인것,이걸로 검색하면 增加检索入口） UF（代） 赠别

赠别 USE（用） 送别

等级关系

等级关系——亦称属分关系，指表示上位概念主题词和下位概念主题词之间的一种关系。

功能： 扩大或缩小检索范围

使用范围——属种关系；整部关系；集合概念与个体概念

哲学

NT（分） 道家

道家

BT（属） 哲学

相关关系

亦称类缘关系，是叙词间除等同关系、等级关系之外语义相关的一种关系。

功能：相关查找。

通常在一词“作为标引词与另一词具有较强的提示作用” 时使用。只在正式主题词之间进行，二者作互逆参照 。

中国共产党第十八次全国代表大会

UF 十八大

RT 中国梦

12-1

主题法核心词汇控制

다른 예시로 설명

主题法어떻게 응용하는지이해

主题法在Web的应用

拼写错误纠正 (mis-spelling correction)

检索词推荐 (query suggestion)

相关搜索词推荐 (related search)

同义词串 (synonym ring)

多义词 (polysemy)

拼写纠正

错误识别：如何判断拼写错误？

Non-word Errors：拼写错误后的词本身就不合法，如错误的将“giraffe”写成“graffe”

Real-word Errors：拼写错误后的词仍然是合法的情况，如将“there”错误拼写为“three” （形近），将“peace”错误拼写为“piece”（同音），将“two”错误拼写为“too”（同音）。

错误纠正：如何找到正确的拼写？

Non-word errors

错误识别 利用词典：任何不被词典所包含的word均被当作spelling error，识别准确率依赖词典的规模和质量

错误纠正 查找词典中与错误拼写最近似（最小编辑距离）的词，例：

## Real-word Errors

25%~40%的拼写错误都属于Real-word类型，与Non-word类型相比，纠错难度更大 错误识别（每个词都是可能的错误对象）

错误纠正：从发音和拼写等角度，查找与word最近似的words集合作为拼写建议，常见的方法有Highest noisy channel probability和Classifier。

检索词推荐

两部分来源

用户输入和点击

类似：买了x的用户还买了y，浏览了a的用户还浏览了b，输入了m的用户还输入了n

词语在网页中的共现频率(co-ocurrence)，例：

计算机 等级考试……90,000个网页中同时出现

计算机 网络…… 80,000

计算机 豆浆机…… 300

相关搜索与检索词推荐的区别

检索词推荐:一般用户进行检索时在检索框里出现，减小用户检索的时间

相关搜索：一般出现在检索结果下面，往往用户的选择检索式的不太好，或检索结果与用户需求不相关的话可以选择。

12-2

广义而言，检索语言包括**受控语言**(**人工语言**)和**自然语言**语言两类语言系统。

狭义而言，检索语言包括规范的受控语言，如**分类表**、**标题表**、**叙词表**、**分类主题**一体化词表等。

检索系统中检索语言的应用变化

基于印刷型文献的手工检索

以受控语言（分类法、主题法）为主

基于数据库的计算机检索 受控语言和自然语言结合

基于网络的信息检索 以自然语言为主

自然语言检索系统

自然语言检索系统就是指对文献作者或文摘提要的编写者原来使用的语言进行一定的序化组织、 处理并提供自然语言检索接口供用户查询使用的 检索系统。

关键性技术：自然语言处理（NLP）

自然语言系统

什么是“自然语言”

选词：在为文献检索标识时，使用文献作者、文摘编写者原来所用的语词或标引人员自拟的语词，而不是取自受控词表中的语词

技术：将自然语言处理技术（NLP）应用于信息系统的信息组织、标引与输出

用户：用自然语言作为提问/查询输入的检索方式

自然语言检索的分类

用户输入检索词形式：关键词检索和自然语言语句的提问式输入检索

检索内容或检索对象：自然语言标引词检索方式和全文检索方式

自然语言处理 Natural Language Processing

自然语言处理（natural language processing,简称NLP）主要研究计算机对输入的自然语言文本的分析、理解和生成。

自然语言处理追求的目标是计算机是如何来理解一个句子和领会 一个文档所要表述的意思。

自然语言处理的核心技术主要是解决自然语言的歧义问题

关键问题：如何建立大规模知识库。

≈ 计算语言学 computational linguistics

自然语言的层次

语音和文字，即基本语言信号的构成；Phonetics & Morphology

词法和句法(合称“语法”)，即语言基本运用单位的构成和组 合的形式规律；Syntax

语义,即语言所要表达的概念结构；Semantics

语用,即语言与语言使用环境的相互作用；Pragmatics

篇章、上下文 Discourse 1

自然语言在信息组织和检索中的应用

中文自动分词

自动标引

自动文摘

全文检索等

自然语言标引

优点

可以降低标引难度及成本，从而提高标引速度；

采用用户熟悉的自然语言，符合用户检索习惯， 减少了概念转换中产生的失真现象，专指度强；

由于自然语言标引检索多采用自动处理方式，省略了编制词表和词汇的智力负担；

操作简单方便，灵活，比较适合没有专业知识的广大网络用户使用等。

缺点

由于自然语言对标引用词不加严格控制，必然会 形成非关键性词语的大量出现，影响检准率的提高；

由于不能反映概念词间的一一对应关系，也不能反映概念关系的隐含性，无法排除同义词（漏检）、近义词、多义词（误检）等词间的模糊现象，势必也会直接影响到检全率的提高。

中文分词/切词

汉语较之于属于屈折型语言的西语系（如英语、法语、 德语等）语言，

词与词之间缺乏任何形式上的标识； 汉语特有的书写形式、灵活多变的构词方式以及不同的分词形式代表着不同含义

分词难点

交集型：如果AB和BC都是词典中的词，那么如果待切分字串中包含“ABC”这个子串，就必然会造成两种可能的切分：“AB/C/” 和 “A/BC/”。

组合型：如果AB和A、B都是词典中的词，那么如果待切分字串中包含"AB" 这个子串，就必然会造成两种可能的切分："AB/" 和"A/ B/ “。

歧义的解决策略

1.排除歧义 常常用词频、词长、词间关系等信息

2.有时切分歧义发生在一小段文字中，但为了排除歧义，需要看较长的一段文字

未登录词识别策略

1. 尽可能多地收集词汇，以降低碰到未登录词的机会；

2. 通过构词规则和上下文特征规则来识别： “EXO单手接奖被指不礼貌 刘德华粉丝会称误会”

3. 通过统计的方法来计算经过一般的分词过程后剩下的“连续单字词碎片”是人名、地名等的可能性，从而识别出未登录词。 “美国犹他州小镇埃斯卡兰特居民艾玛坤斯乐

自然语言标引

索引词的选择范围

人工索引－>质量高，但不适用大规模文档数据处理

自动索引

部分索引－>title, abstract, keywords, etc（例如：书目 系统）

全文索引－>文档中所有词 都参与索引 (例如：Web Search Engine)

索引词的选择原则

Index term ≠ word

理想：表达文档内容的语义单位

字、词、短语

自然语言标引

方式

自由标引（人工）

自动标引

1.自动抽词标引

2.自动赋词标引

3.自动赋分类号标引

4.单汉字索引与全文索引

自动标引、自由标引 优缺点

自动抽词标引이 있어야 나머지 과정도 가능함

自由标引

自由标引即人工关键词标引

是指人工对表征文献主题内容具有实质意义的语词，亦即对揭示和描述文献主题内容来说是重要的、关键性的（可以作为检索入口的）语词抽取，抽取时不加规范或只做少量规范处理

自由标引的优点

由于不使用词表控制，标引速度要比使用词表的主题标引快、标引成本低；

可用与文献主题专指度一致的词进行标引，保证较高的检准率；

标引过程是标引人员进行主题分析的过程，如果标引人员具有一定的专业水平，则其标引质量可大大高于自动抽词标引

自动标引

自动标引（Automatic Indexing），是根据文献内容，依靠计算机系统全部或部分地自动给出标引符号的过程

换句话说，是利用计算机系统模仿人的标引活动并自动生成检索所需的索引符号的过程

自动标引的分类

自动主题标引

自动分类标引

自动主题标引关键： 词语抽取；对所识别出的主题词进行优选，以最终产生能表达文献主题内容的标引词。

自动标引的优点

1. 适应信息资源快速增长的需要

2. 处理能力强、速度快、成本低

3. 一致性好 Cleverton的实验：两位有经验的标引员用同一 叙词表对同一篇文献进行标引，其标引词的 同一率仅有30%左右

4. 自动标引不受标引人员状态和情绪的影响，稳定性好

抽取关键词的主要方法

绝对词频统计法：以词在文章中出现的绝对频次为根本依据确定文章的中心关键词，理论基础是Zipf’s Law齐普夫定律。

词频权重法：除考虑词频外，还考虑词的位置、词的词性、 词本身的价值、词的长度等因素，对词进行加权，然后根据 权值大小确定关键词。

Zipf’s law

齐普夫定律对于自动抽词标引说明了什么？

高频词：没有实际意义，例如 the, of, and

低频词：无法作为表达文献主题的词语

理想的索引词：中频词

13-2

全文检索技术

所谓全文检索（Full-Text Retrieval），是指以全部文本信息作为 检索对象的一种信息检索技术。全文检索的核心技术就是维护一 个高效索引（主要是倒排文档）。

다 봐야할 듯

14-1

网络信息资源组织方式.

1.主题树方式（Open Directory、雅虎分类目录）

2.文件组织方式（FTP）

3.数据库组织方式（Deep Web，搜索引擎无法索引）

4.超文本组织方式 (hyper-text) 通过节点和超链来实现对信息的组织

基于主题（自然语言索引和检索技术）

基于超链分析（以PageRank为代表的一系列排序 算法，HITS-Hubs and Authorities）

网页的链接

网页的链接结构主要是指各网页间的链接形式， 一般分为：

树状结构（层次结构）

网状结构（平级结构）

只依赖TF\*IDF有什么问题？

这里的相关性主要是基于内容的相关性但是会有一些垃圾网页，虽然也包含大量的查询词，但却并非满 足用户需要的文档,

简化版PageRank的问题

所有的Rank值都来自于链接

没有链接的网页Rank值为零（하이퍼링크를 통하지 않고 검색창을 통해서 들어오는 경우를 포함하지못함）

Pagerank算法相关概念

PageRank值（PR值） 用来评价网页的重要性，PR值越大越重要

阻尼系数d：(damping factor，一般取0.85) 定义：用户不断随机点击链接的概率，它取决于点击的次数，被设定为0-1之间。

d的值越高，继续点击链接的概率就越大。

用户停止点击并随机访问另一页面的概率在式子中用常数(1-d)表示

Danggling node

所有只有入链、没有出链的网页的rank值在下一次递归时均分给所有网页

网页的Rank值，由Dangling Matrix和Hypertext Matrix共同构成

14-2

웹1.0 2.0 차이와 공통점

联系

通过Web,网络上的资源，可以在一个网页中比较直观的表示出来

资源之间，在网页上可以互相链结。

差异

1.Web1.0是单向网络，web2.0是双向互动的网络，用户既是网络信息资源的生产者，又是信息资源的消费者。

Web 2.0

Web 2.0是让网络用户从信息接受者转变成为信息制造者和传播者、从受众转向主体、从个体转向社团的新型互联网服务模式。

特征：

平等交互：信息主体与受众、信息消费者与生产者的地位是等同的。

社会化：强调用户之间的社会化协作。

个性化：将互联网建成以用户为中心的网络，它的价值和意义在于为用户带来了真正的个性化和信息自主权

Web 2.0技术与应用

1. 用于发布个人信息的平台，带有聚合性质。代表有各式各样的博客 （Blog）、个人空间（如MSN Space）、RSS (站点摘要)等。

2. 社交网路服务（Social Network Service, SNS），根据「 六度分隔理论」 ， 通过网际网路技术手段实现的社交网络服务，代表有Linkedin、 Friendster、 openBC等。

3. 具有 Tag (标签)功能的计算机协同工作软件，在某些软件工程师群体中极其流行。

4. 由用户合作完成某种任务或服务的平台。例如世界各地用户同时编辑的 Wiki(维基百科全书)，Mechanic Turk等

标签

互联网内容组织方式

用户用于描述内容的关键字，以便于检索和分享。

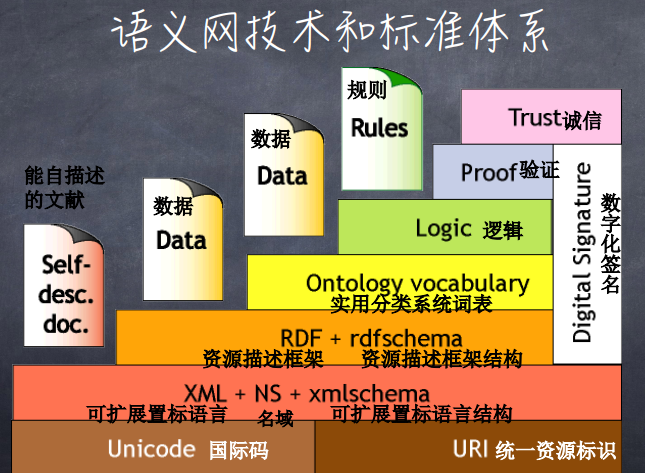
标签将内容的组织权利从网站管理者下放到用户手中，充分体现了web2.0自下而上，用户参与的特点。

分析实例，用到哪些信息组织的方式，话题的本质是标签

15-1

Web 3.0 그림

매층의 기능



Unicode和URI层

Unicode和URI是整个语义Web的基础 Unicode处理资源的编码，保证使用的是国际通用字符集，实现 网上信息的统一编码。

URI（统一资源标识）是URL（Universal Resource Locator-即网 址）的超集，URI支持语义网上的对象和资源的精细标识，从而使精确信息检索成为可能。

Xml rdf 语义网

XML+命名空间+XML模式层

XML层具有命名空间（Name Space）和XML Schema（XML模式）

定义 通过XML标记语言将网上资源信息的结构、内容与数据的表现形式进行分离，确保语义网的定义，并支持与其他基于XML的标准进行无缝集成。

RDF중요

RDF 资源描述框架（Resource Description Framework），资源描述框架 资源描述框架（RDF）是由W3C开发的一个资源描述规范。其最初的目标就是解决不同元数据的互操作问题 。RDF是一个用于表达关于Web资源的元数据，比如Web页面的标题、作者和修改时间，Web文档的版权和许可信息，某个被共享资源的可用计划表等。

RDF与XML

1．RDF使用了XML语言，RDF在XML基础之上，以一种标准的、能够相互操作的方式揭示数据的深层语义

2．RDF和XML是相互补充的。RDF主要关注于元数据模型的建立， 而有关编码方式、字符集等等方面的问题，RDF依赖于XML

3．还应当了解到，基于XML的RDF仅仅只是RDF的一种应用模式， 除XML语法之外，RDF还允许有其它的语法模式

RDF基本思想

1．用Web标识符（称作统一资源标识符，Uniform Resource Identifiers或URI）来标识事物

2．用简单的属性（property）及属性值来描述资源

3．这使得RDF用资源标识-属性类型-属性值这样的表示模型来定义一个或多个关于资源

RDF+RDF Schema层

该层用于描述语义网上的资源及其类型为网上资源描述提供了一种通用框架和实现数据集成的元数据解决方案最底层的URI标识网上的对象，RDF和RDFS层则可对URI标识的对象进行陈述（Statement1

本体层

1.用于描述各种资源之间的联系

2．揭示资源本身以及资源之间更为复杂和丰富的语义信息

3．将信息的结构和内容相分离，对信息作完全形式化的描述，使网上信息具有计算机可理解的语义。

4 因为本体定义了不同概念间的关系，所以本体层能够对词汇 （Vocabularies）的变迁提供支持。

逻辑层、验证层

逻辑主要提供公理和推理规则，为智能推理提供基础。该层用来产生规则。

验证（Proof）证明注重于提供认证机制，证明层执行逻辑层产生的规则，并结合信任层的应用机制来评判是否能够信赖给定的证明。

诚信（Trust）

顶层的诚信注重于提供信任机制，以保证用户代理Agent在网上进行个性化服务和彼此间交互合作时，更安全与可靠

15-2

需求的差别

移动端的需求

**移动端的需求都是基于特定场景**

**浏览减小、对个性化信息的依赖增加**

移动端的需求应用场景

信息检索：**简单事实**、**答案**、**而不是需要较长阅读时间的内容**

大量的信息需求与出行相关：地点，时间，天气等

个人信息管理

策略나 장단점은 안외워도됨

桌面段的区别