第九章：

文档自动摘要技术

杨建武

北京大学计算机科学技术研究所

Email:yangjianwu@icst.pku.edu.cn

文摘的定义

- 以提供文献内容梗概为目的，不加评论和补充解释，简明、确切地记述文献重要内容的短文。**(GB6447-86文摘编写规则)**
- An express of a certain document without any explanations and comment. It's unnecessary to know who writes the summary. (**ANSI**)
- A concise and accurate express of the document without any explanation and comment. A summary is independent on the author of the summary. (**ISO214-1976(E)**)
- Concise(简洁)，Accurate(准确)，Explicit(清楚)

文摘的种类(GB6447—86)

- 报道性文摘 informative abstracts
  - 指明文献的主题范围及内容梗概的简明文摘，也称简介。
- 报道性/指示性文摘 informative-indicative abstracts
  - 以报道性文摘的形式表述文献中信息价值较高的部分，而以指示性文摘的形式表述其余部分的文摘。
- 作者文摘 author’s abstracts
  - 由文献作者自己撰写的文摘。
- 文摘员文摘 abstractpr’s abstracts
  - 由文献作者以外的人员编写的文摘。

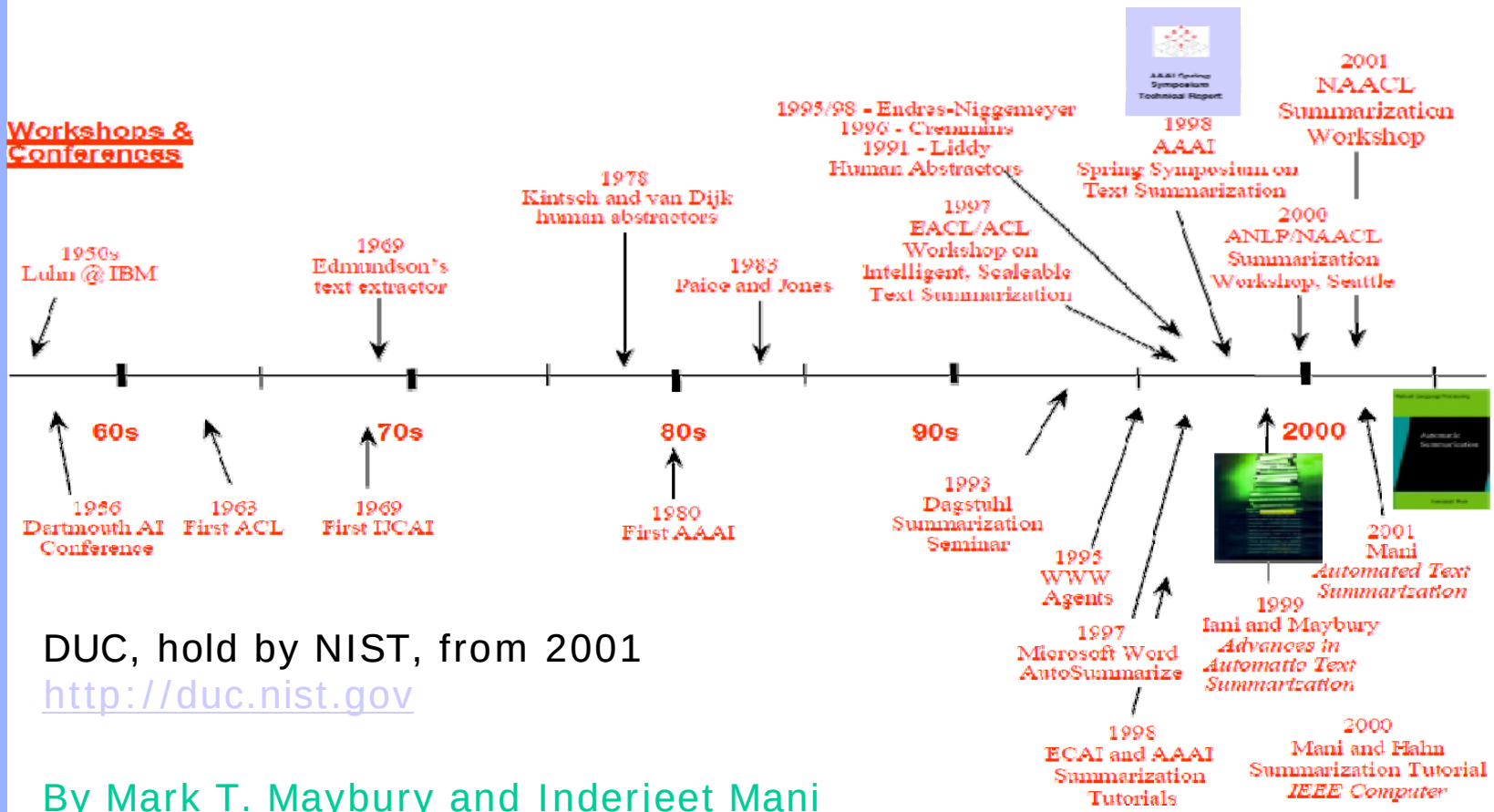
Summary Classification

- Classified by user's requirement
  - Generic Summarization (GS)
  - User-query Summarization (UQS)
- Classified by text object
  - Single Document Summarization
  - Multiple Document Summarization
- Classified by method
  - Summarization Based on Extraction (SBE)
  - Summarization Based on Understanding (SBU)
- Classified by need corpus
  - Supervised Summarization (SS)
  - Unsupervised Summarization (US)

自动摘要

- 定义：
  - 利用计算机自动地从原始文档中提取全面准确地反映该文档中心内容的简单连贯的短文。
- 自动文摘系统
  - 自动文摘系统应能将原文的主题思想或中心内容自动提取出来。
  - 文摘应具有概况性、客观性、可理解性和可读性。
  - 系统适用于任意领域。

A Brief History of Summarization



DUC, hold by NIST, from 2001
http://duc.nist.gov

By Mark T. Maybury and Inderjeet Mani

研究现状

- 国外研究主要是面对英文信息的处理，比较有代表性的系统有：
  - 美国哥伦比亚大学的多文档自动文摘系统***Newsblaster***。
    - 对每天发生的同主题新闻进行摘要。
  - 美国密西根大学研究开发的***WebInEssence***
    - 个性化的基于Web的多文档自动文摘和内容推荐系统。
  - 美国南加利福尼亚大学的信息科学研究所***NeATS***。
  - Vivisimo公司 （http://www.vivisimo.com）
  - infonetware公司（http://www.infonetware.com）
    - 这两个公司对搜索引擎返回的结果进行了有效地聚类整理。
    - 文档聚类是多文档自动文摘的一个关键的预处理步骤。
- http://www.summarization.com/
- DUC (Document Understanding Conference)
- 北大、中科院、哈工大、复旦、上海交大等

评价方法

Evaluation

- 内部评价方法(Intrinsic Methods)
  - 在提供参考摘要的前提下，以参考摘要为基准评价系统摘要的质量。通常情况下，系统摘要与参考摘要越吻合，其质量越高。
- 外部评价方法(Extrinsic Methods)
  - 不需要提供参考摘要，利用文档摘要代替原文档执行某个文档相关的应用。
  - 例如文档检索、文档聚类、文档分类等，能够提高应用性能的摘要被认为是质量好的摘要。例如在搜索

Evaluation-- Edmundson

- Edmundson评价
  - 客观评估：比较机械文摘（自动文摘系统得到的文摘）与目标文摘的句子重合率(coselection rate)。
  - 主观评估：由专家比较机械文摘与目标文摘所含的信息，然后给机械文摘一个等级评分。等级分为：完全不相似，基本相似，很相似，完全相似等。

Evaluation-- Edmundson

- Edmundson评价的几个基本规定：
  - 专家文摘和机械文摘都存入文本文件中；
  - 比较的基本单位是句子；
    - 句子是两个句子级标点符号之间的部分。
    - 句子级标号包括：“。”“：”“；”“！”“？”；
  - 为使专家文摘与机械文摘具有可比性，只允许专家从原文中抽取句子，而不允许专家根据自己对原文的理解重新生成句子；
  - 专家文摘和机械文摘的句子都按照在原文中出现的先后顺序给出。

Evaluation

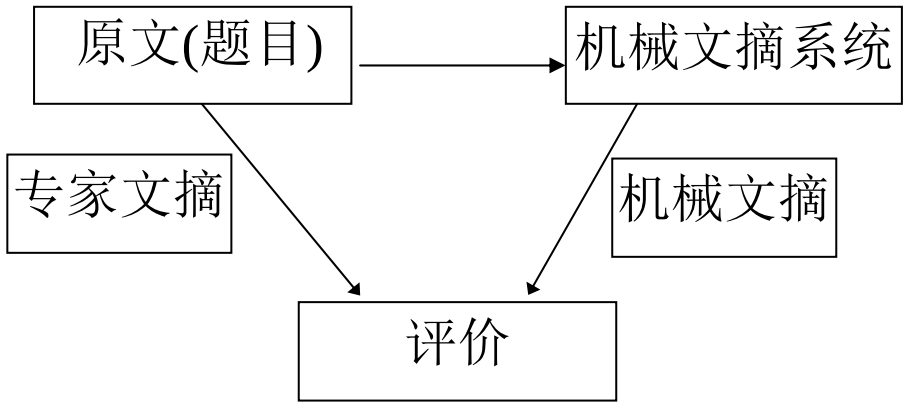
- 定义：

重合率p＝匹配句子数 / 专家文摘句子数×100％

- 每一个机械文摘的重合率为按三个专家给出的文摘得到的重合率的平均值。

平均重合率＝ $\sum_{i=1}^{n} P_i / n * 100\%$

（Pi为相对于第i个专家的重合率，n为专家的数目）



$$\mathrm{Re}call = N_{hm} / N_h$$

$$\mathrm{Pr}ecision = N_{hm} / N_m$$

Evaluation--ROUGE准则

- 由ISI的Lin和Hovy提出的一种自动摘要评价方法
- 被广泛应用于DUC的摘要评测任务中
- ROUGE准则基于摘要中n元词(n-gram)的共现信息来评价摘要，是一种面向n元词召回率的评价方法
- ROUGE准则由一系列的评价方法组成，包括：
  - ROUGE-1，ROUGE-2，ROUGE-3，ROUGE-4
    - （其中ROUGE-1至ROUGE-4分别基于1元词到4元词）
  - 以及ROUGE-L，ROUGE-W等

Evaluation--ROUGE准则

$$ROUGE-N=\frac{\sum_{S\in\{Ref\ Summaries\}}\sum_{n\text{-}gram\in S}Count_{match}(n-gram)}{\sum_{S\in\{Ref\ Summaries\}}\sum_{n\text{-}gram\in S}Count(n-gram)}$$

- 其中：
  - n-gram表示n元词，
  - {Ref Summaries}表示参考摘要，
  - $Count_{match}$(n-gram)表示系统摘要和参考摘要中同时出现n-gram的个数，
  - Count(n-gram)则表示参考摘要中出现的n-gram个数。

DUC

- http://duc.nist.gov/
- The Document Understanding Conference (DUC) is a series of summarization evaluations that have been conducted by the National Institute of Standards and Technology (NIST) since 2001.
- Its goal is to further progress in automatic text summarization and enable researchers to participate in large-scale experiments in both the development and evaluation of summarization systems.
- Since 2008, DUC has moved to the Text Analysis Conference (TAC) http://www.nist.gov/tac/
  - Question Answering; Recognizing Textual Entailment; Summarization

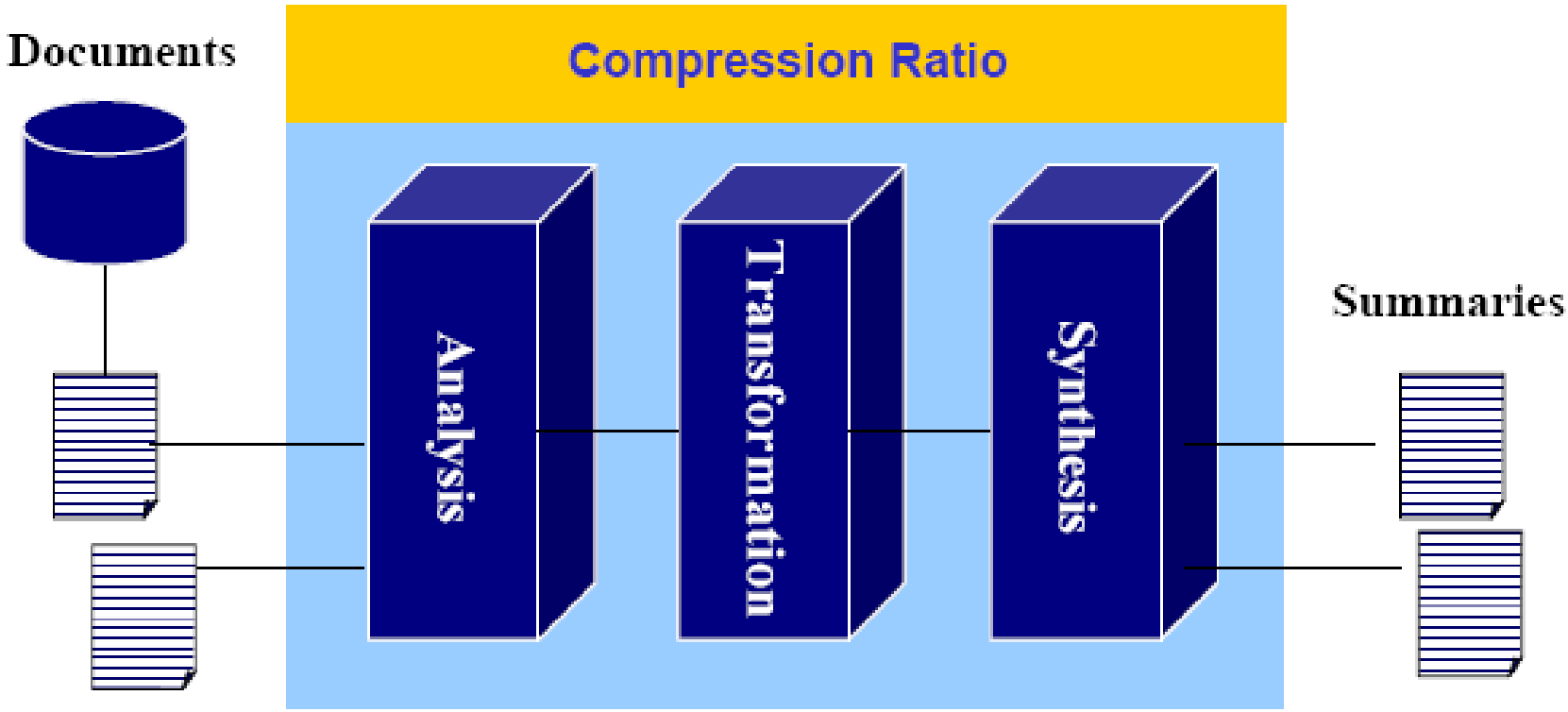
自动摘要方法

Summarization Algorithms

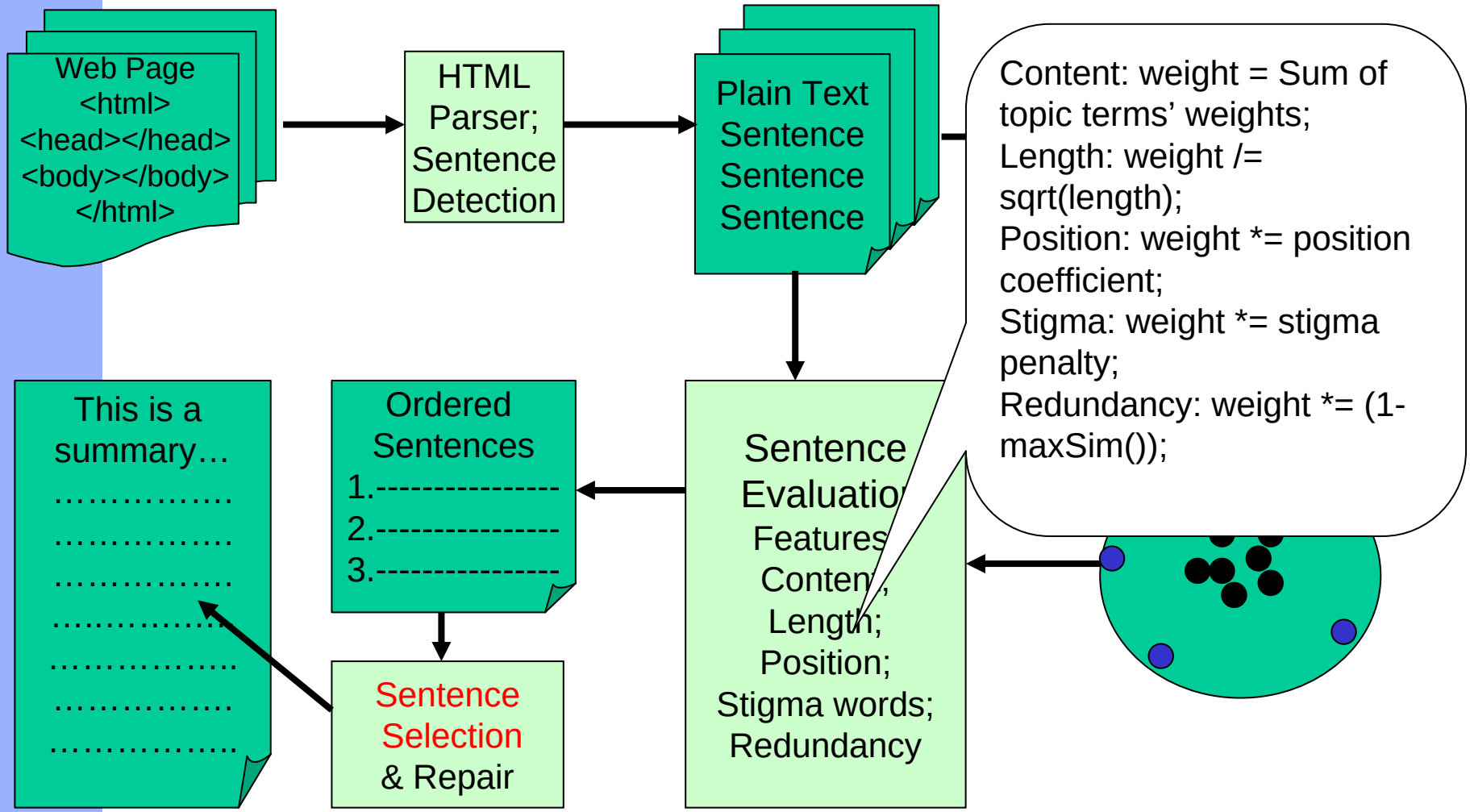
- Keyword summaries
  - Display most significant keywords
  - Easy to do
  - Hard to read, poor representation of content
- Sentence extraction
  - Extract key sentences
  - Medium hard
  - Summaries often don't read well
  - Good representation of content
- Natural language understanding / generation
  - Build knowledge representation of text
  - Generate sentences summarizing content
  - Hard to do well

Something between the last two methods?

自动摘要


Documents
Compression Ratio
Analysis
Transformation
Synthesis
Summaries

Summarization Review



基本方法

位置法

- 美国的P.E.Baxendale的研究结果显示：人工摘要中的句子为段首句的比例为85%，是段尾句的比例为7%。
- 美国康奈尔大学G.Salton提出了寻找文章的中心段落为文摘核心的思想。
- 其他
  - E.g.： 除了论题句、段首、段尾等句子之外，段落的第二句常常表示段落的主题。

提示字串法

- 文章中常常有一些特殊的线索词(短语、字串、字串链)，它们对文章主题具有明显的提示作用，可以利用它们来获取文章的主题。
- e.g: Edmundson的文摘系统中的线索词词典:
  - 取正值的奖励词(Bonus Words)
  - 取负值的惩罚词(Stigma Words)
  - 无效词(Null Words)

频率统计法

- 实验表明：高频字串往往与主题相关度极大。
- [Luhn,1958]：根据句子中实词的个数来计算句子的权值。
- [V.A.Oswald] 主张句子的权值应按其所含代表性的“词串”的数量来计算；
- [Doyle]则重视共现频度最高的“词对”；
- [Lisa.F.Rau,1995]采用相对词频的方法实现ANES(Autormatic News Extraction System)系统。

文章框架法

- 目次性摘要：借助文章的大小标题与语义段的摘要方法。
- 统计表明：大部分科技文献(99.8%)的标题都能基本反映主题。
- 捷克Janos把文中的句子分为主干句与枝叶句，删枝叶句留主干句的文摘方法可划归于“文章框架法”。

信息提取法

- 信息提取法常用于对一些特殊领域的文献资料做摘要(如气象预报等)。
- 该方法根据用户的需求，
  - 首先构造出一个用户喜闻乐见的文摘框架(Abstract Frame)，文摘框架以空槽的形式提出应该从原文中获取的各项内容，
  - 然后再把文摘框架中的内容转换为文摘(文字或图表)。
- 该方法常称之为二段式：抽取有关信息，然后生成摘要。

理解分析法

- 基于理解的自动摘要常包含语法分析、语义分析、信息提取和文摘生成，作者文摘应属于此。
- 研究表明：理解首先应着重篇章理解、段落理解，也就是理解应该是分层的，高层理解比低层理解更为重要。

仿人算法

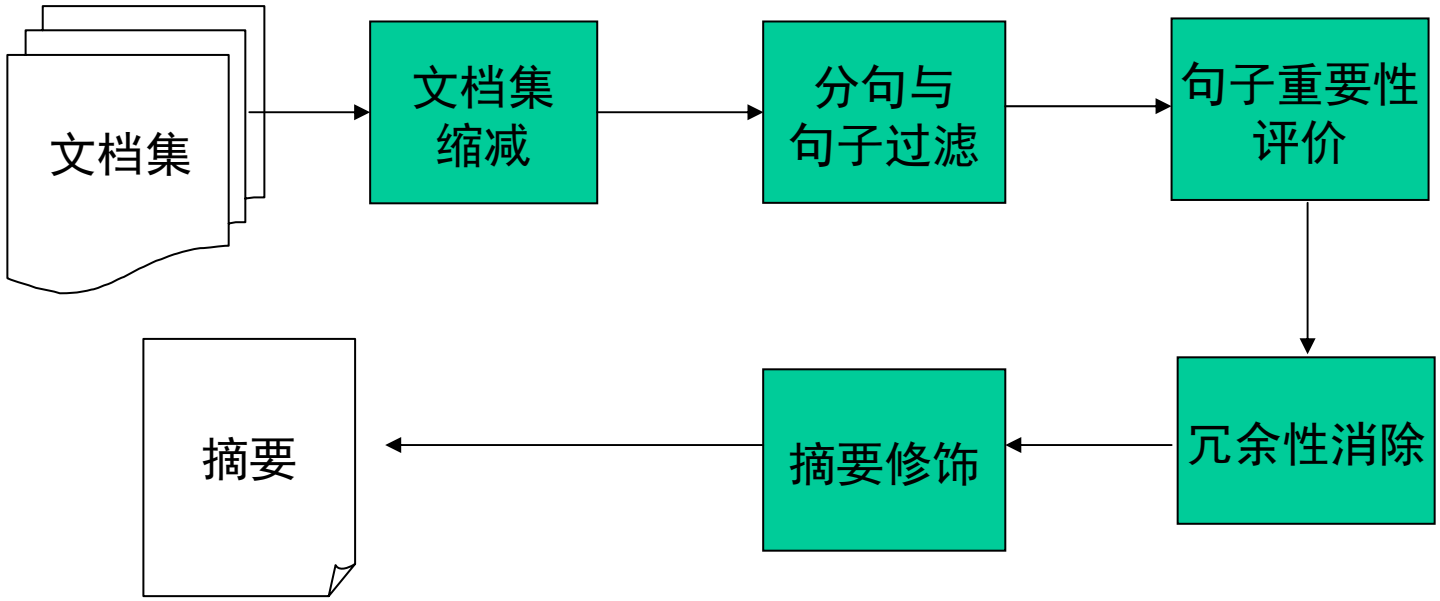
- 仿人算法就是对人工方法的学习，模仿与发挥所产生的综合性方法。
- 手工文摘人员在编制文摘时并不一定通读全文，往往只着重观察标题、前言、结束语及其论题句，以发现其主题，再挑选句子并修饰稍加组织生成文摘。
- 人工很多经验都是值得注意的，同一篇文献，不同用户兴趣点和观察角度可能不同，文摘的结果应当不同。

Sentence Extraction

Sentence Extraction

- Represent each sentence as a feature vector
- Compute score based on features
- Select n highest-ranking sentences
- Present in order in which they occur in text.
- Postprocessing to make summary more readable/concise
  - Eliminate redundant sentences
  - Anaphors/pronouns (代词)
  - Delete subordinate clauses, parentheticals (插入语)

基本方法



Sentence Importance--Score

- Statistical features
  - Tf
  - Tf-idf
- Linguistic features
  - Location
  - Semantic
- Integrative features

$$Score(S_i) = \lambda * \sum_{s \in S} w_s * (Q_s . S_i) + (1 - \lambda) * \sum_{l \in L} w_l * (L_l . S_i)$$

Document Structure Analysis

- Paragraph analysis
  - 段落的位置
  - 句子在段落中的位置
- Sentence boundary analysis
  - 句子级标号
  - 分句后的指代消歧

消除冗余

- 冗余性消除:
  - 类似MMR：根据摘要中的句子消除待选择句子的冗余性:
  - 对于待选择句子s_i,

    $score(s_i)=score(s_i)-sim(s_i,s_j)*score(s_j)$;

    其中s_j为摘要中与s_i最相似的句子。

摘要修饰

- 摘要修饰：
  - 目的：对句子进行排列，使生成的摘要保持好的连贯性和可读性；
  - 排序、指代消解等

A Trainable Document Summarizer

- Sigir95 paper on summarization by Kupiec, Pedersen, Chen
- Trainable sentence extraction

Feature Representation

- Fixed-phrase feature
  - Certain phrases indicate summary, e.g. “in summary”
- Paragraph feature
  - Paragraph initial/final more likely to be important.
- Thematic word feature
  - Repetition is an indicator of importance
- Uppercase word feature
  - Uppercase often indicates named entities. (Taylor)
- Sentence length cut-off
  - Summary sentence should be > 5 words.

Training

- Hand-label sentences in training set (good/bad summary sentences)
- Train classifier to distinguish good/bad summary sentences
- Model used: Naïve Bayes

$$P(s \in S | F_1, F_2, \ldots F_k) = \frac{\prod_{j=1}^{k} P(F_j | s \in S)\ P(s \in S)}{\prod_{j=1}^{k} P(F_j)}$$

- Can rank sentences according to score and show top n to user.

Evaluation of features

- Baseline (choose first n sentences): 24%
- Overall performance (42-44%) not very good.
- However, there is more than one good

Feature	Individual Sents Correct	Cumulative Sents Correct
Paragraph	163 (33%)	163 (33%)
Fixed Phrases	145 (29%)	209 (42%)
Length Cut-off	121 (24%)	217 (44%)
Thematic Word	101 (20%)	209 (42%)
Uppercase Word	100 (20%)	211 (42%)

Clustering based Algorithm

Clustering Algorithm

- 理论依据：语篇语言学的理论认为，语篇在意义上存在一种层次关系，即：
  - 语篇的中心意思 ＝各组成意义段的中心意思按一定逻辑关系的组合
  - 意义段的中心意思 ＝各组成子意义段的中心意思按一定逻辑关系的组合
  - 子意义段的中心意思 ＝各组成下位子意义段的中心意思按一定逻辑关系的组合
  - 直至不能再划分为更小的子意义段。

Sentences clustering

- 利用自动聚类将文档分为若干段落类
- 从中选出与文档主题相关的段落类作为候选段落类
- 最后从候选段落类中选出句子构成摘要

面向主题的文档摘要及MMR Algorithm

Query-Specific Summarization

- A **generic summary** makes no assumption about the reader's interests.
- **Query-specific summaries** are specialized for a single information need, the query.
- Summarization is much easier if we have a description of what the user wants.
- Recall from last quarter:
  - Google-type excerpts – simply show keywords in context

MMR Algorithm

- Maximal Marginal Relevance
- MEAD：http://www.summarization.com/mead/
- 方法：
  - 在选择文摘句时，使要进入文摘的句子既和主题的相关度较高，又使该句和已选文摘句之间的冗余度尽可能的小
  - 从而保证句子和主题或用户Query的相关，同时减少冗余信息，增加有特色的内容，使得到的文摘质量较高。

$$MMR \equiv \text{Arg} \max_{D_i \in R \setminus A} [\lambda \text{Sim1}(Di,Q) - (1-\lambda) \max_{D_j \in A} \text{Sim2}(D_i, D_j)]$$

MMR相关方法

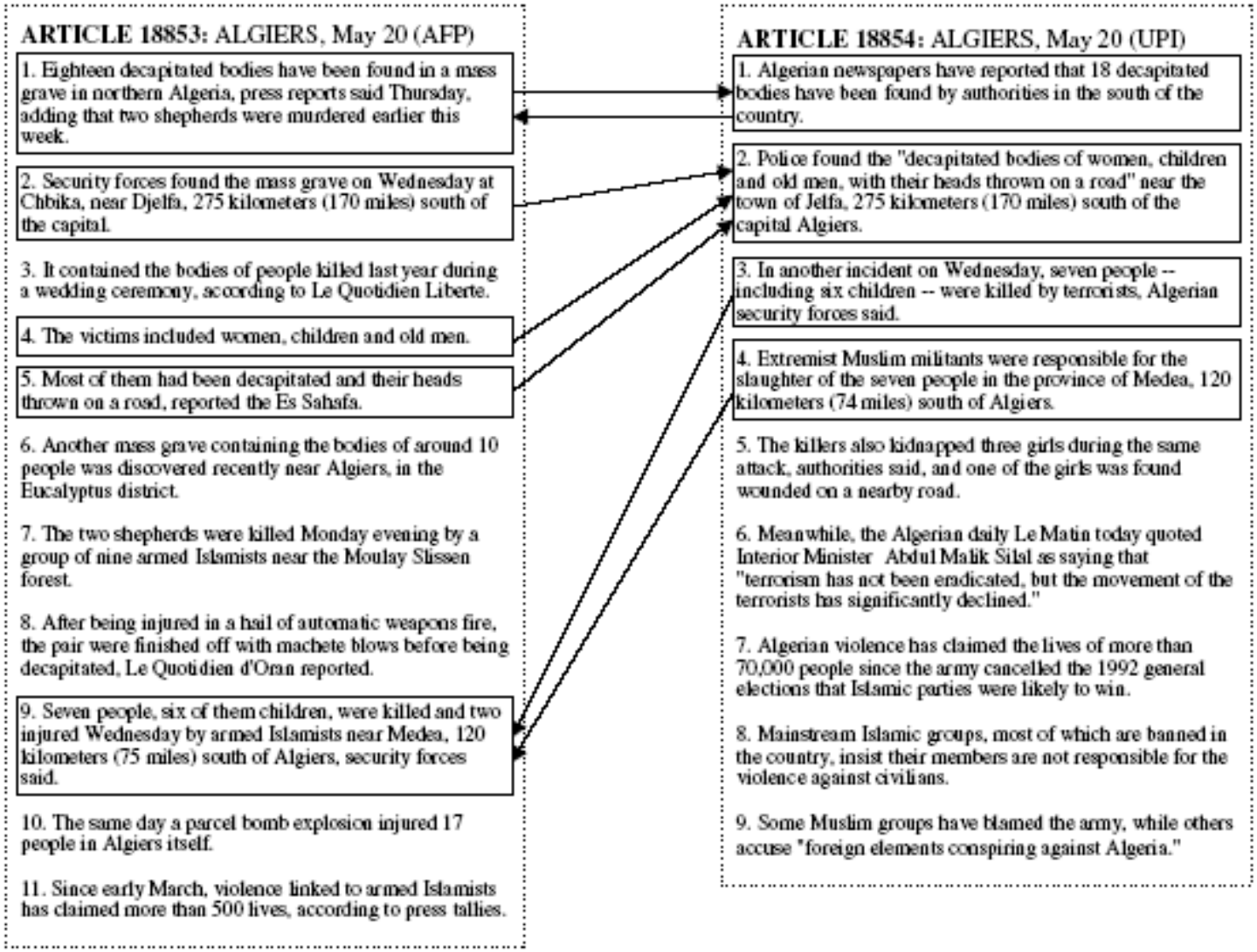
- MMI-MS
  - 日本横滨国立大学开发的一个多文档自动文摘系统将MMR技术和IGR (Information Gain Ratio)技术结合起来，称为MMI-MS (Maximal Marginal Importance – Multi-Sentence)来选取文摘句。
- MMR-MD
  - Goldstein等提出了在多文档文摘系统中采用基于MMR-MD (Maximal Marginal Relevance Multi-Document)的方法。
- MMR-SS
  - 哈工大刘寒磊,关毅等提出了基于句子语义相似的最大边缘相关方法:MMR-SS (Semantic Similarity based Maximal Marginal Relevance)来选择文摘句，生成关于同一主题的通用文摘。

多文档摘要

多文档摘要

- multi-document summarization
- 意义：
  - 服务于主题检测、聚类等模块；为文档集生成简明描述，方便用户浏览，辅助用户决策；

难点


ARTICLE 18853: ALGIERS, May 20 (AFP)
1. Eighteen decapitated bodies have been found in a mass grave in northern Algeria, press reports said Thursday, adding that two shepherds were murdered earlier this week.
2. Security forces found the mass grave on Wednesday at Chbika, near Djelfa, 275 kilometers (170 miles) south of the capital.
3. It contained the bodies of people killed last year during a wedding ceremony, according to Le Quotidien Liberte.
4. The victims included women, children and old men.
5. Most of them had been decapitated and their heads thrown on a road, reported the Es Sahafa.
6. Another mass grave containing the bodies of around 10 people was discovered recently near Algiers, in the Eucalyptus district.
7. The two shepherds were killed Monday evening by a group of nine armed Islamists near the Moulay Slissen forest.
8. After being injured in a hail of automatic weapons fire, the pair were finished off with machete blows before being decapitated, Le Quotidien d'Oran reported.
9. Seven people, six of them children, were killed and two injured Wednesday by armed Islamists near Medea, 120 kilometers (75 miles) south of Algiers, security forces said.
10. The same day a parcel bomb explosion injured 17 people in Algiers itself.
11. Since early March, violence linked to armed Islamists has claimed more than 500 lives, according to press tallies.
ARTICLE 18854: ALGIERS, May 20 (UPI)
1. Algerian newspapers have reported that 18 decapitated bodies have been found by authorities in the south of the country.
2. Police found the "decapitated bodies of women, children and old men, with their heads thrown on a road" near the town of Jelfa, 275 kilometers (170 miles) south of the capital Algiers.
3. In another incident on Wednesday, seven people -- including six children -- were killed by terrorists, Algerian security forces said.
4. Extremist Muslim militants were responsible for the slaughter of the seven people in the province of Medea, 120 kilometers (74 miles) south of Algiers.
5. The killers also kidnapped three girls during the same attack, authorities said, and one of the girls was found wounded on a nearby road.
6. Meanwhile, the Algerian daily Le Matin today quoted Interior Minister Abdul Malik Silal as saying that "terrorism has not been eradicated, but the movement of the terrorists has significantly declined."
7. Algerian violence has claimed the lives of more than 70,000 people since the army cancelled the 1992 general elections that Islamic parties were likely to win.
8. Mainstream Islamic groups, most of which are banned in the country, insist their members are not responsible for the violence against civilians.
9. Some Muslim groups have blamed the army, while others accuse "foreign elements conspiring against Algeria."

Types of MD Summaries

- Single event/person tracked over a long time period
  - Elizabeth Taylor's bout with pneumonia
  - Give extra weight to character/event
  - May need to include **outcome (dates!)**
- Multiple events of a similar nature
  - Marathon runners and races
  - More broad brush, ignore dates
- An issue with related events
  - Gun control
  - Identify key concepts and select sentences accordingly

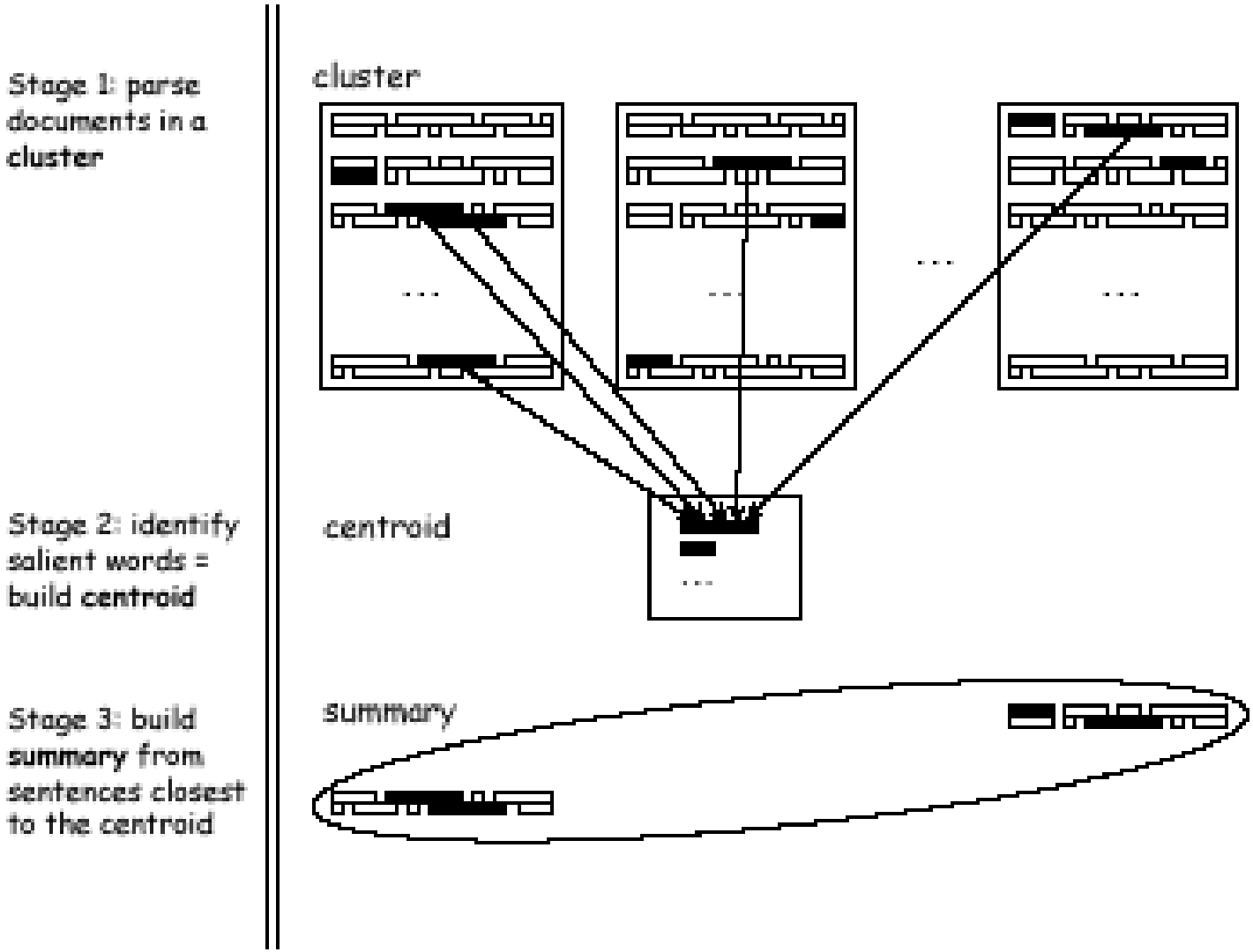
Determine MD Summary Type

- First, determine which type of summary to generate
- Compute all pairwise similarities
- Very dissimilar articles → multi-event (marathon)
- Mostly similar articles
  - Is most frequent concept named entity?
  - Yes → single event/person (Taylor)
  - No → issue with related events (gun control)

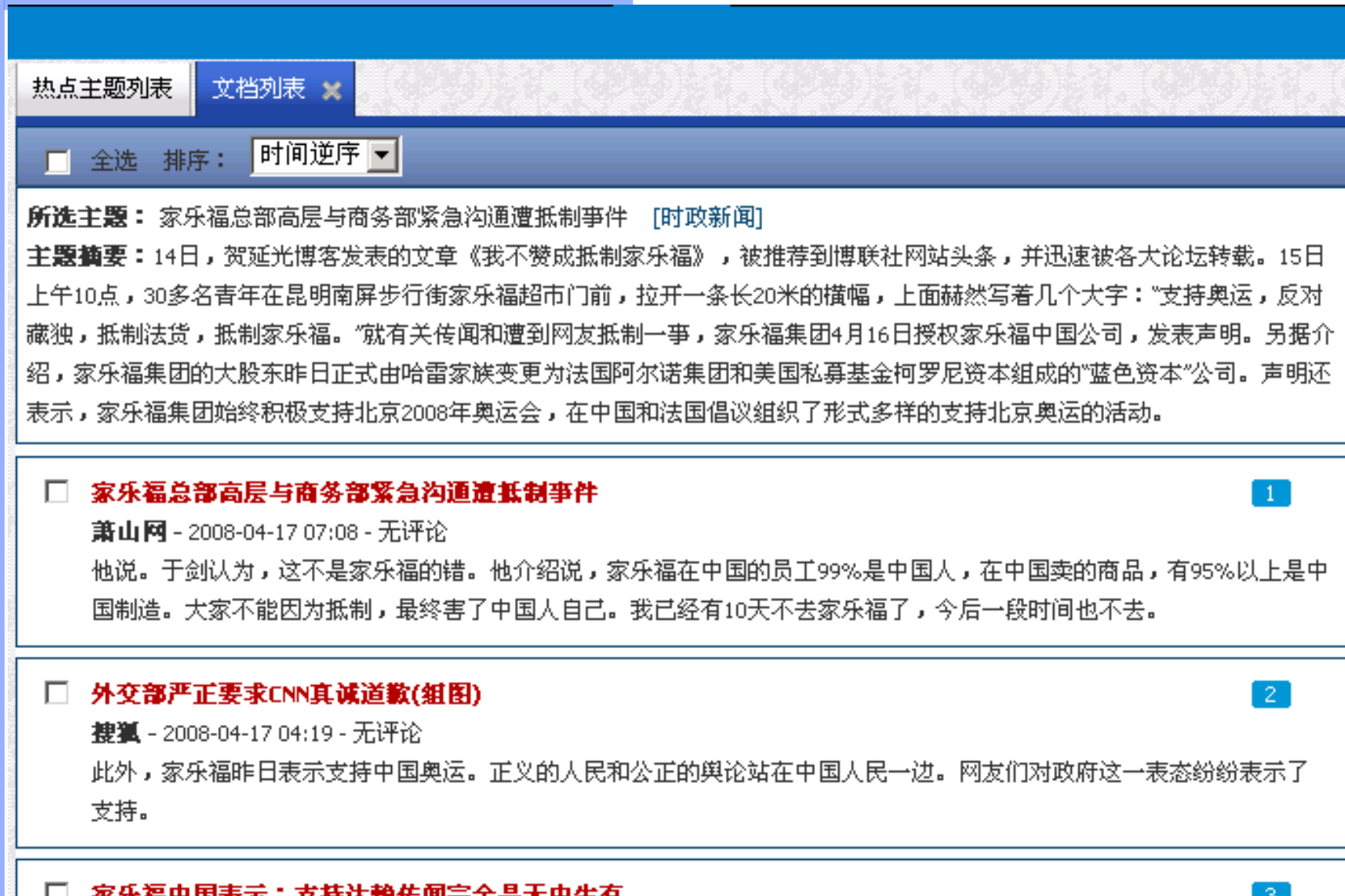
Centroid-based summarization (MEAD)

- 首先对句子进行聚类
- 然后综合考虑句子级特征以及句子之间的特征来对句子进行重要性评价，
  - 特征包括类簇中心点值、句子位置、与首句的重叠度等。
- 其中类簇中心点值表示一个句子包含的中心词的权重之和，这个值越大，说明该句子越重要；
- 句子位置也反映了句子的重要性，
  - 一个句子在文章中越靠前，那么这个句子越重要；
    - 文章首句通常很重要这个前提

Centroid-based summarization (MEAD)



应用实例


热点主题列表
文档列表
全选
排序：
时间逆序
所选主题：家乐福总部高层与商务部紧急沟通遭抵制事件 [时政新闻]
主题摘要：14日，贺延光博客发表的文章《我不赞成抵制家乐福》，被推荐到博联社网站头条，并迅速被各大论坛转载。15日上午10点，30多名青年在昆明南屏步行街家乐福超市门前，拉开一条长20米的横幅，上面赫然写着几个大字："支持奥运，反对藏独，抵制法货，抵制家乐福。"就有关传闻和遭到网友抵制一事，家乐福集团4月16日授权家乐福中国公司，发表声明。另据介绍，家乐福集团的大股东昨日正式由哈雷家族变更为法国阿尔诺集团和美国私募基金柯罗尼资本组成的"蓝色资本"公司。声明还表示，家乐福集团始终积极支持北京2008年奥运会，在中国和法国倡议组织了形式多样的支持北京奥运的活动。
家乐福总部高层与商务部紧急沟通遭抵制事件
1
萧山网 - 2008-04-17 07:08 - 无评论
他说。于剑认为，这不是家乐福的错。他介绍说，家乐福在中国的员工99%是中国人，在中国卖的商品，有95%以上是中国制造。大家不能因为抵制，最终害了中国人自己。我已经有10天不去家乐福了，今后一段时间也不去。
外交部严正要求CNN真诚道歉(组图)
2
搜狐 - 2008-04-17 04:19 - 无评论
此外，家乐福昨日表示支持中国奥运。正义的人民和公正的舆论站在中国人民一边。网友们对政府这一表态纷纷表示了支持。

小结

- 文档摘要的概念
- 文档摘要的评价
- 基本方法
- 面向主题的文档摘要
- 多文档摘要

Any Question?