自动文档摘要(续)，文本分类与文本聚类

社交网络与舆情预测 第四讲

任昭春

Outline

- Clustering-based document summarization (con’t)
- 文档摘要的最新进展
- 文本分类
- 文本聚类

The basic MRW model (3/3)

- Limitations of current MRW model:
  - 对文档集合内所有句子同等对待和使用
  - 没有考虑子主题类簇的高层次信息
- How to incorporate the cluster-level information into the process of sentence ranking?
  - **Cluster-based Conditional Markov Random Walk Model**
  - **Cluster-based HITS Model**

The proposed models (1/7)

- Three steps:
  - Theme cluster detection;
  - Sentence score computation;
  - Summary extraction.

The proposed models (2/7)

- Theme cluster detection
  - Kmeans Clustering
  - Agglomerative Clustering
  - Divisive Clustering

$$k = \sqrt{|V|}$$

The proposed models (3/7)

- Cluster-based Conditional Markov Random Walk Model
  - Incorporates the cluster-level information into the link graph
  - Based on PageRank;
  - $G^*=<V_s, V_c, E_{ss}, E_{sc}>$
  - $V_s=V=\{v_i\}$; $V_c=C=\{c_j\}$
  - $E_{ss}=E=\{e_{ij}|v_i,v_j \in V_s\}$
  - $E_{sc}=\{e_{ij}|v_i \in V_s, c_j \in V_c$ and $c_j=clus(v_i)\}$

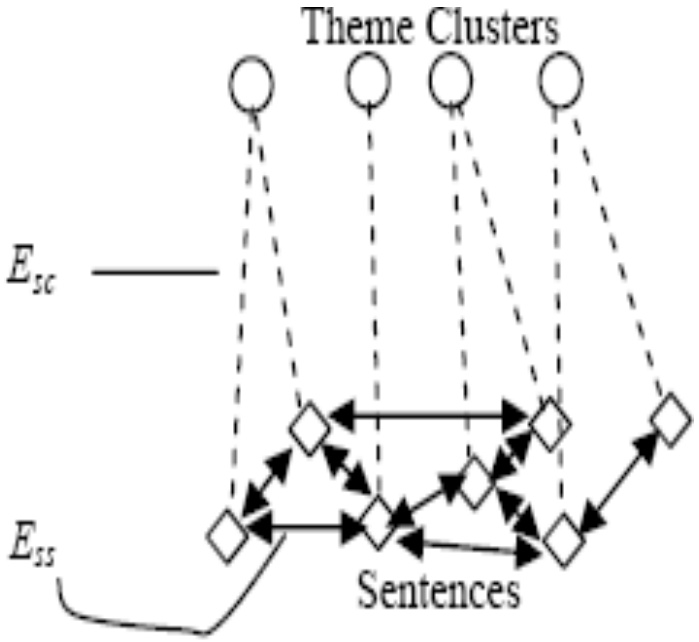


Figure 2: Two-layer link graph

The proposed models (4/7)

- Cluster-based Conditional Markov Random Walk Model

$$p(i \to j \mid clus(v_i), clus(v_j)) = \begin{cases} \dfrac{f(i \to j \mid clus(v_i), clus(v_j))}{\sum_{k=1}^{|V|} f(i \to k \mid clus(v_i), clus(v_k))}, & \text{if } \sum f \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

$$f(i \to j \mid clus(v_i), clus(v_j))$$

$$= \lambda \cdot f(i \to j \mid clus(v_i)) + (1-\lambda) \cdot f(i \to j \mid clus(v_j))$$

$$= \lambda \cdot f(i \to j) \cdot \pi(clus(v_i)) \cdot \omega(v_i, clus(v_i))$$

$$+ (1-\lambda) \cdot f(i \to j) \cdot \pi(clus(v_j)) \cdot \omega(v_j, clus(v_j))$$

$$= f(i \to j) \cdot (\lambda \cdot \pi(clus(v_i)) \cdot \omega(v_i, clus(v_i))$$

$$+ (1-\lambda) \cdot \pi(clus(v_j)) \cdot \omega(v_j, clus(v_j)))$$

The proposed models (5/7)

- Cluster-based Conditional Markov Random Walk Model

$$\pi(clus(v_i)) = sim_{\cos ine}(clus(v_i), D)$$

$$\omega(v_i, clus(v_i)) = sim_{\cos ine}(v_i, clus(v_i))$$

$$\tilde{M}^*_{i,j} = p(i \rightarrow j \mid clus(v_i), clus(v_j))$$

$$A^* = \mu \tilde{M}^{*T} + \frac{(1-\mu)}{|V|} \vec{e}\vec{e}^T$$

The proposed models (6/7)

- Cluster-based HITS Model
  - Considers the clusters and sentences as hubs and authorities in the HITS algorithm
  - Based on HITS;
  - $G^{\#}=<V_s, V_c, E_{sc}>$
  - $V_s=V=\{v_i\}$ $V_c=C=\{c_j\}$
  - $E_{sc}=\{e_{ij}|v_i \in V_s, c_j \in V_c\}$

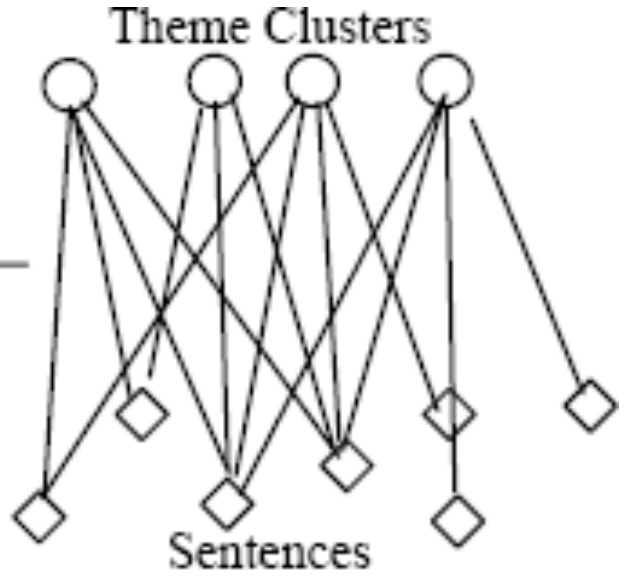


Figure 3: Bipartite link graph

The proposed models (7/7)

- Cluster-based HITS Model

$$L_{i,j} = w_{ij} = sim_{\cos ine}(v_i, c_j)$$

$$AuthScore^{(t+1)}(v_i) = \sum_{c_j \in V_c} w_{ij} \cdot HubScore^{(t)}(c_j) \qquad \vec{a}^{(t+1)} = L\vec{h}^{(t)}$$

$$HubScore^{(t+1)}(c_j) = \sum_{v_i \in V_s} w_{ij} \cdot AuthScore^{(t)}(v_i) \qquad \vec{h}^{(t+1)} = L^T \vec{a}^{(t)}$$

$$\vec{a}^{(t+1)} = \vec{a}^{(t+1)} / \left|\vec{a}^{(t+1)}\right| \qquad \vec{h}^{(t+1)} = \vec{h}^{(t+1)} / \left|\vec{h}^{(t+1)}\right|$$

Experiments and Results (1/4)

- Datasets

Table 1: Summary of data sets

	DUC 2001	DUC 2002
Task	Task 2	Task 2
Number of documents	309	567
Number of clusters	30	59
Data source	TREC-9	TREC-9
Summary length	100 words	100 words

- Evaluation Metrics
  - ROUGE-1, ROUGE-2, ROUGE-W

Experiments and Results (2/4)

Table 2: Comparison res

System	ROUGE-1
ClusterCMRW (Kmeans)	0.35824
ClusterCMRW (Agglomerative)	0.35707
ClusterCMRW (Divisive)	0.35549
ClusterHITS (Kmeans)	0.35756
ClusterHITS (Agglomerative)	0.36897*
ClusterHITS (Divisive)	0.37419*
MRW	0.35527
SystemN	0.33910
SystemP	0.33332
SystemT	0.33029
Coverage	0.33130
Lead	0.29419

Table 3: Comparison results on DUC2002

System	ROUGE-1	ROUGE-2	ROUGE-W
ClusterCMRW (Kmeans)	0.38221*	0.08321	0.12362
ClusterCMRW (Agglomerative)	0.38546*	0.08652*	0.12490*
ClusterCMRW (Divisive)	0.37999	0.08389	0.12384*
ClusterHITS (Kmeans)	0.37643	0.08135	0.12141
ClusterHITS (Agglomerative)	0.37768	0.07791	0.12271
ClusterHITS (Divisive)	0.37872	0.08133	0.12282
MRW	0.37595	0.08304	0.12173
System26	0.35151	0.07642	0.11448
System19	0.34504	0.07936	0.11332
System28	0.34355	0.07521	0.10956
Coverage	0.32894	0.07148	0.10847
Lead	0.28684	0.05283	0.09525

Experiments and Results (3/4)

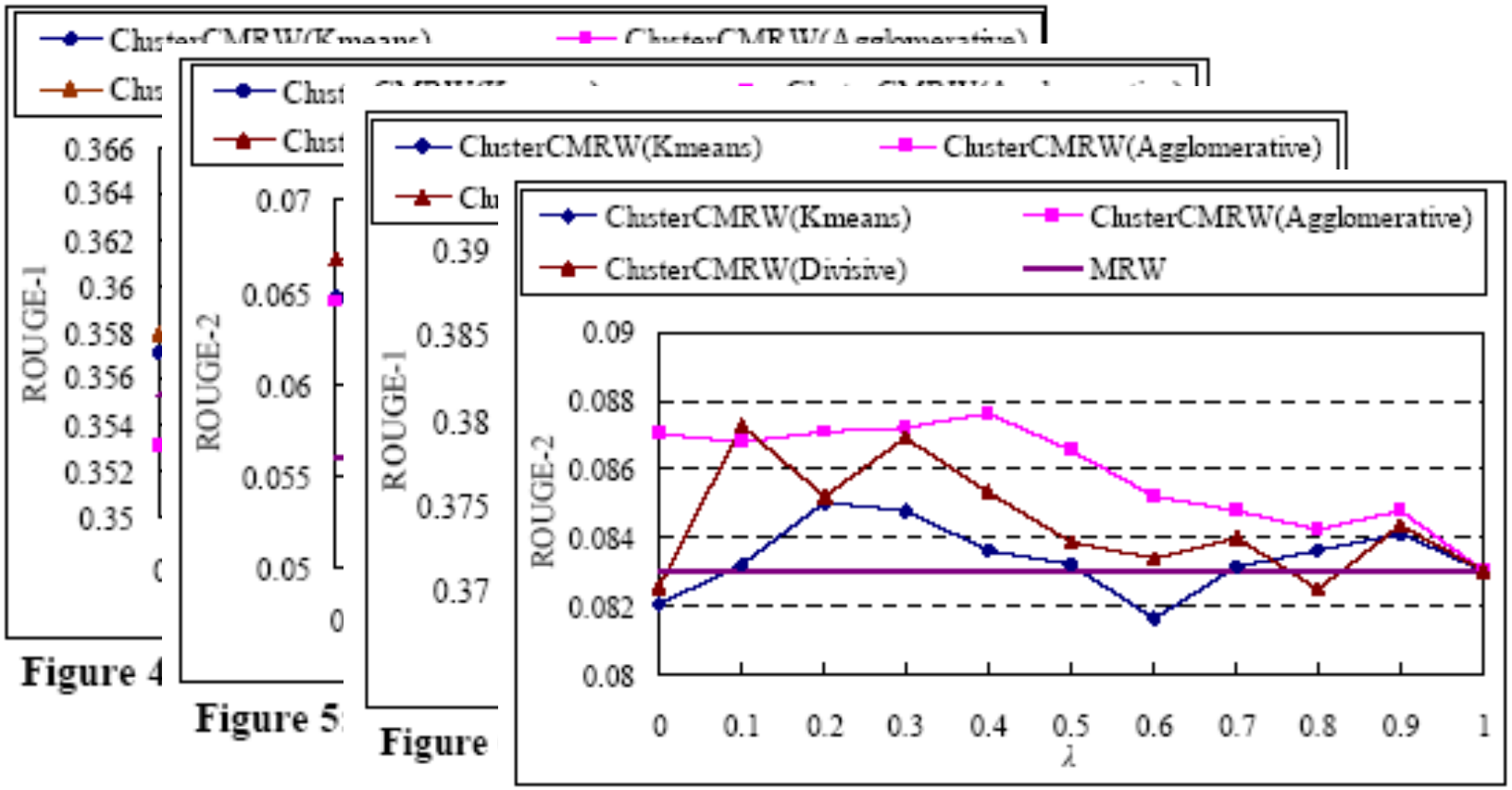


Figure 4

Figure 5

Figure

Figure 7: ROUGE-2 vs. λ for ClusterCMRW on DUC2002

Experiments and Results (4/4)

- ClusterCMRW vs. ClusterHITS

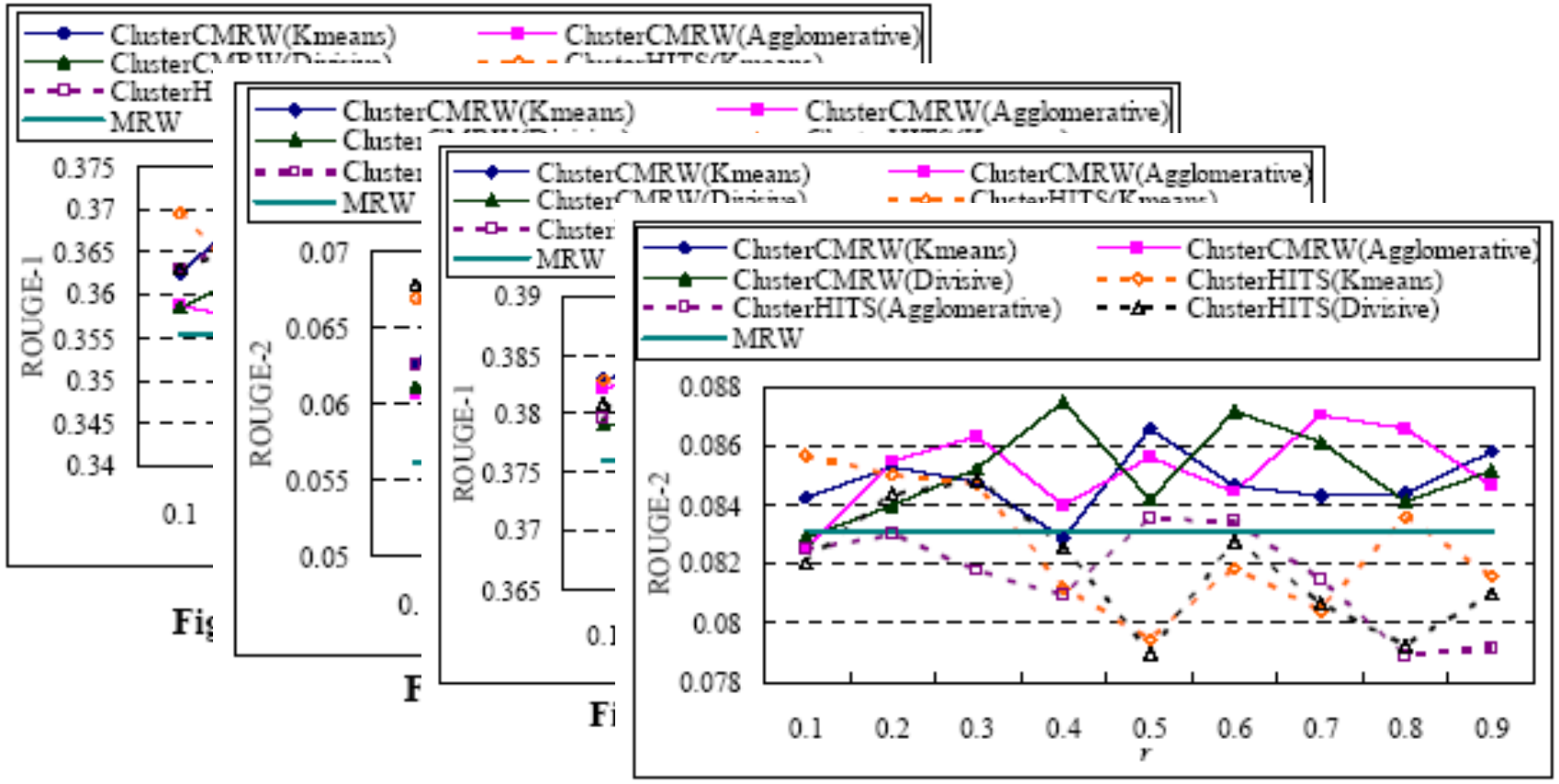


Figure 11: ROUGE-2 vs. *r* on DUC2002

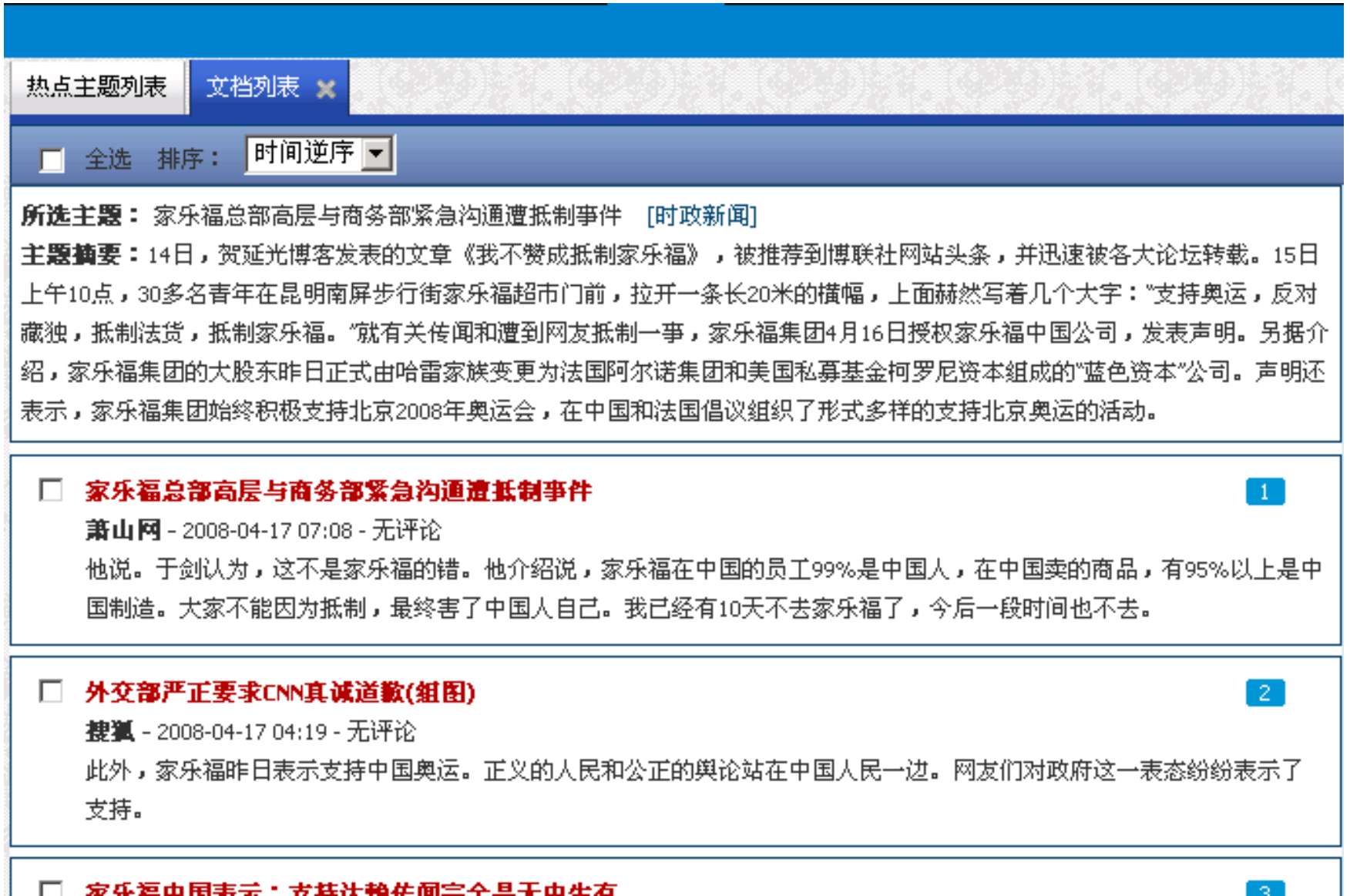
Conclusions

- 提出的两个模型都比较有效;
- ClusterCMRW比ClusterHITS鲁棒性更好;

相比较PageRank，HITS计算效率较低，这是因为我们不可以离线计算HITS值，必须要在每次查询时计算一次。除此之外，主题漂移(Topic drift)也是HITS算法存在的一大问题。

HITS算法每次都在一个子图中运算，导致容易受到低质量链接影响

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。

对特殊类型文档摘要

- 科技文献摘要
- 电子邮件摘要
- 网页、网站摘要
- 书籍摘要
- 多媒体摘要：视音频摘要
- …

对文档摘要的新应用需求

- 查询相关的多文档摘要 (SIGIR08；ACL06)
- “更新”式摘要 (DUC07,DUC08)
- 超短摘要：标题自动生成(COLING02)
- 综述文章自动生成 (IJCAI99)
- 移动终端(PDA,手机)上的文档摘要(WWW2001)
- 情感摘要(ACL08)
- 人物传记式摘要(ACL08)
- 演化式摘要(SIGIR04)
- 比较式摘要
- …

小结

- 文档摘要的概念
- 文档摘要的评价
- 基本方法
  - Sentence Extraction
  - A Trainable Document Summarizer
  - 面向主题的摘要(MMR Algorithm)
- 多文档摘要
  - Centroid-based summarization (MEAD)
  - CollabSum: Exploiting Multiple Document Clustering for Collaborative Single Document Summarizations (SIGIR 2007)
  - Multi-Document Summarization Using Cluster-Based Link Analysis (SIGIR 2008)

Outline

- Clustering-based document summarization (con’t)
- 文档摘要的最新进展
- 文本分类
- 文本聚类

在传统文档摘要上的研究

- **探索新闻摘要中的新线索**
  - **计算句子确定性并用于新闻摘要**

Lincoln High School, on the edge of Portland's downtown, has cut all music programs.

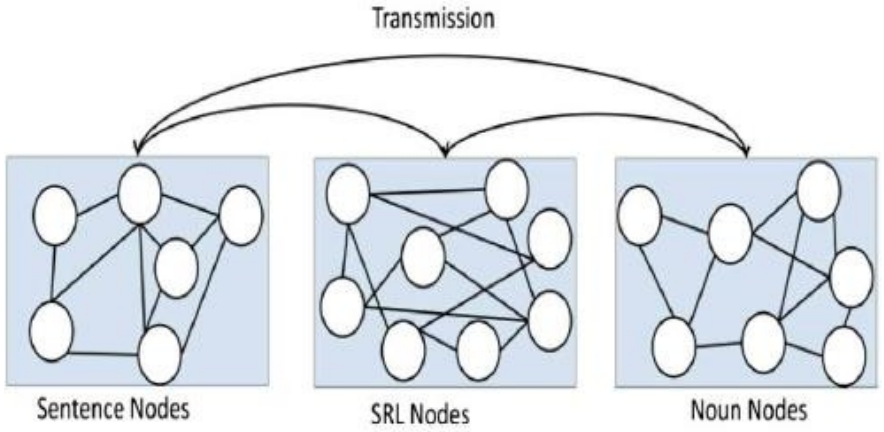
However, it *seems* that Obama will not use the platform to relaunch his stalled drive for Israeli-Palestinian peace.

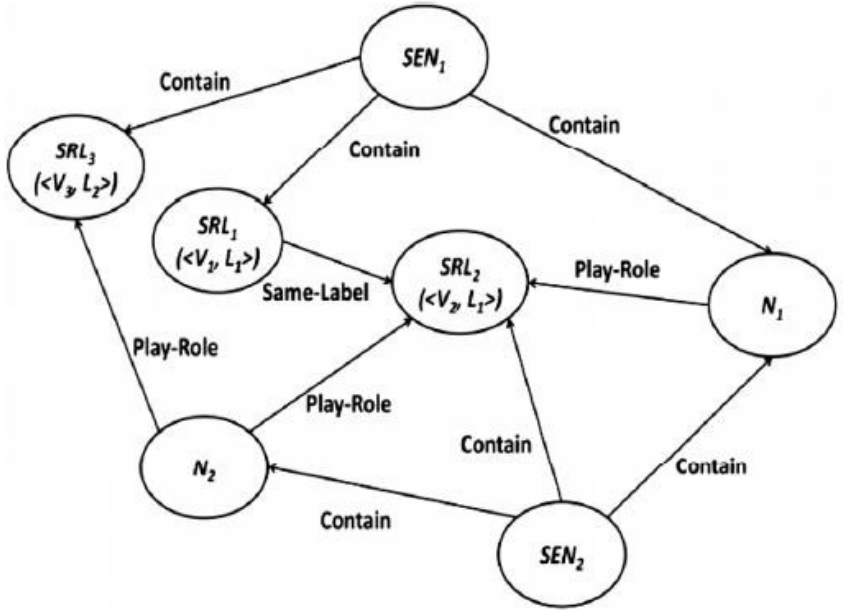
$$f^{new}(s_i, s_j) = f(s_i, s_j) \times (1 + \lambda \cdot CertainScore(s_j))$$
$$= sim_{cosine}(s_i, s_j) \times (1 + \lambda \cdot CertainScore(s_j))$$

Xiaojun Wan and Jianmin Zhang. "CTSUM: Extracting More Certain Summaries for News Articles." SIGIR-2014.

在传统文档摘要上的研究

- **有效利用句法语义信息**
  - **将语义角色标注信息引入图排序模型**

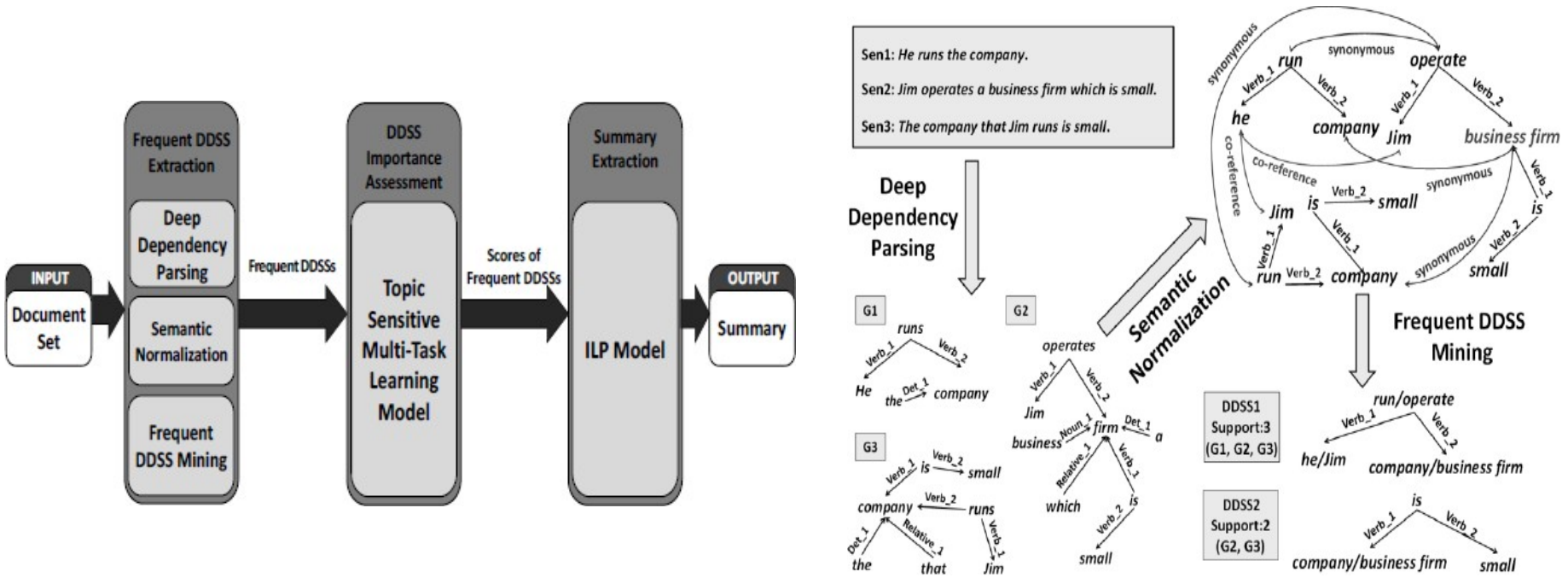




Su Yan and Xiaojun Wan. "SRRank: Leveraging Semantic Roles for Extractive Multi-Document Summarization." IEEE TASLP, 2014.

在传统文档摘要上的研究

- **有效利用句法语义信息**
  - **利用深层依存分析结果进行重要性学习与句子选择**



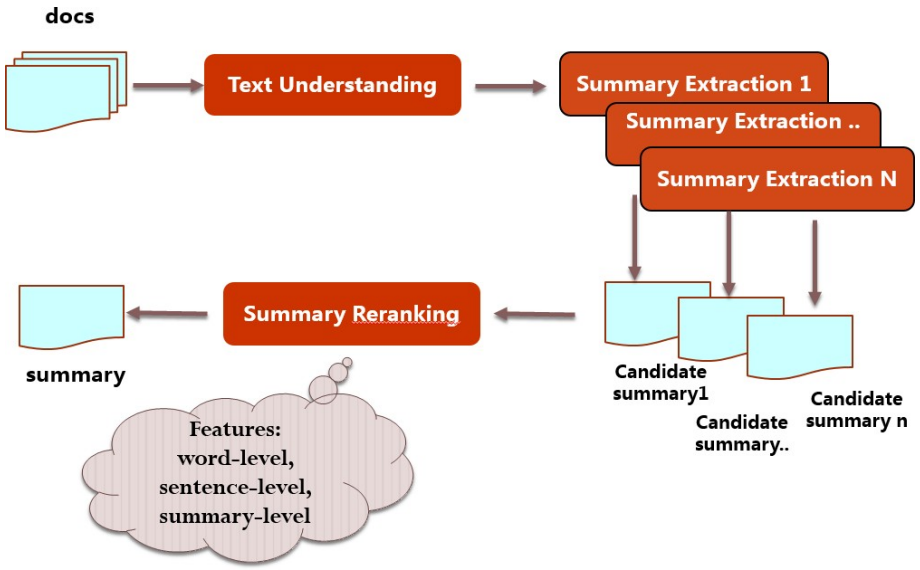
Su Yan and Xiaojun Wan. "Deep Dependency Sub-Structure Based Learning for Multi-Document Summarization." ACM TOIS,2015.

在传统文档摘要上的研究

- **基于稀疏优化的压缩式摘要**

$$\min_{R,A} \quad \|D - RA\|^2 + \lambda_1 \|A\|_{2,1} + \lambda_2 \sum_{i=1}^{n} \|R_i\|_1$$
$$s.t. \quad r_{ij}, a_{ij} \geq 0, \ \forall i, j; \ grammatical(R_i).$$

  - **基于数据重构思想**
- **基于摘要排序的多文档摘要**
  - **候选摘要抽取：为每个文档集抽取多个候选摘要**
  - **摘要排序：从多个候选摘要中选择最优摘要**



Jin-ge Yao, Xiaojun Wan and Jianguo Xiao. "Compressive Document Summarization via Sparse Optimization." IJCAI-2015.

Xiaojun Wan, Ziqiang Cao, Furu Wei, et al. "Multi-Document Summarization via Discriminative Summary Reranking." arXiv preprint arXiv:1507.02062 (2015).

面向互联网大数据的文本摘要研究

- **大数据的3V共性**
  - **规模巨大**
  - **类型多样**
    - **新闻、网页、论文、电子邮件、书籍、微博等等**
  - **时效性强**

- **文本大数据的其他特性**
  - **多语言性**
  - **观点性**
  - **…**

上述特性给信息摘要技术带来了全新的挑战与机遇。

面向文本大数据的新型摘要技术

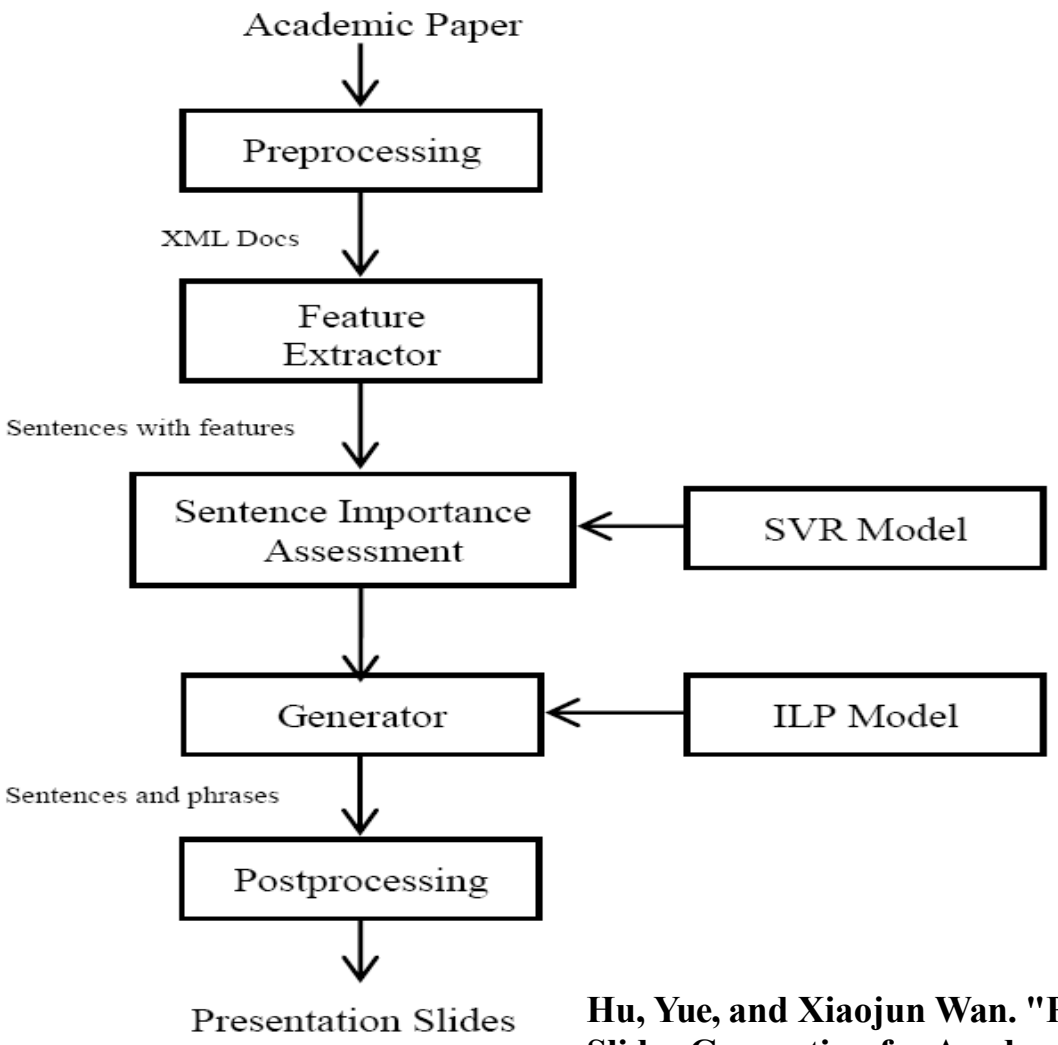
- **大数据的3V共性**
  - **规模巨大 => 增量式摘要**
  - **类型多样 => 针对特定类型文本的摘要**
  - **时效性强 => 更新式摘要、演化式摘要**

- **文本大数据的其他特性**
  - **多语言性 => 跨语言摘要**
  - **观点性 => 观点摘要、比较式摘要**

重点

- **学术文献摘要与文稿生成**
  - 学术文献有良好的篇章结构，有引用关系
  - 可定义新颖的摘要任务
- **跨语言摘要**
  - 针对机器翻译效果的不理想
  - 方便读者快速了解其它语言撰写的新闻/文档内容

学术文献摘要之演示文稿的自动生成



Hu, Yue, and Xiaojun Wan. "PPSGen: Learning-Based Presentation Slides Generation for Academic Papers." TKDE2015.

A Framework for Monitoring Agent-Based Normative Systems

1 INTRODUCTION

- Agent Behaviours
  - In this paper we describe a generic framework for monitoring of agent behaviours in normative multi-agent systems.
  - Observers of agent behaviours are explicitly entrusted by the system's participating agents to accurately report on these behaviours.
- Enforcement Approach
  - In contrast, the enforcement approach [1, 2, 6, 14, 18] allows for autonomous agents, and hence the possibility of violation of norms by agents.
  - The enforcement approach thus requires that agent actions are monitored; that is, they must be observable and recognised as complying with or violating norms, in order that the enforcement mechanisms be appropriately applied.

1 INTRODUCTION

- Enforcement Mechanisms
  - Enforcement mechanisms are thus required to motivate agent compliance Cite as: A Framework for Monitoring Agent-Based Normative Systems, S.Modgil, N.
  - We propose a trusted observer model with observations of agent messages and states of interest, to provide some measure of assurance that enforcement mechanisms are appropriately applied, so encouraging deployment of agents in normative systems.

1 INTRODUCTION

- Monitor Agents
  - This paper also describes how individual norms — obligations, prohibitions, and permissions — can be represented as Augmented Transition Networks (ATNs) [21] that are processed by monitor agents, together with observations relayed to the monitors by trusted observers, in order to determine the fulfilment and violation status of norms.
  - In Section 4, we describe how individual norms are represented as ATNs, and processed by monitor agents.
  - We report on a proof of concept implementation of a monitoring agent, and its processing of ATN representations of norms encoded in an electronic contract specified by the CONTRACT project 1 .

2 A GENERAL MODEL OF NORMS

- Normexpiration
  - Finally N 's NormExpiration denotes the state of interest under which the norm is no longer in force.
  - Henceforth, we will refer to a norm's NormActivation, Norm Condition and NormExpiration as a norm's components.

3 TRUSTED OBSERVERS AND THE MONITORING ARCHITECTURE (The Monitoring Architecture)

- Atn Representations
  - A mapper maps norms to their ATN representations, for input to the monitor (this input is provided off-line).
  - The monitor can identify which observers to subscribe to, based on the ATN representation (as will be described in Section 4).
  - Monitors process observations together with the ATN representations of the norms, to determine when a norm is activated, fulfilled or violated, or has expired.
- Observers And Monitors
  - The behaviours of observers (and monitors) may themselves be governed by normative clauses, and thus observed and monitored for deviation from their expected behaviour.

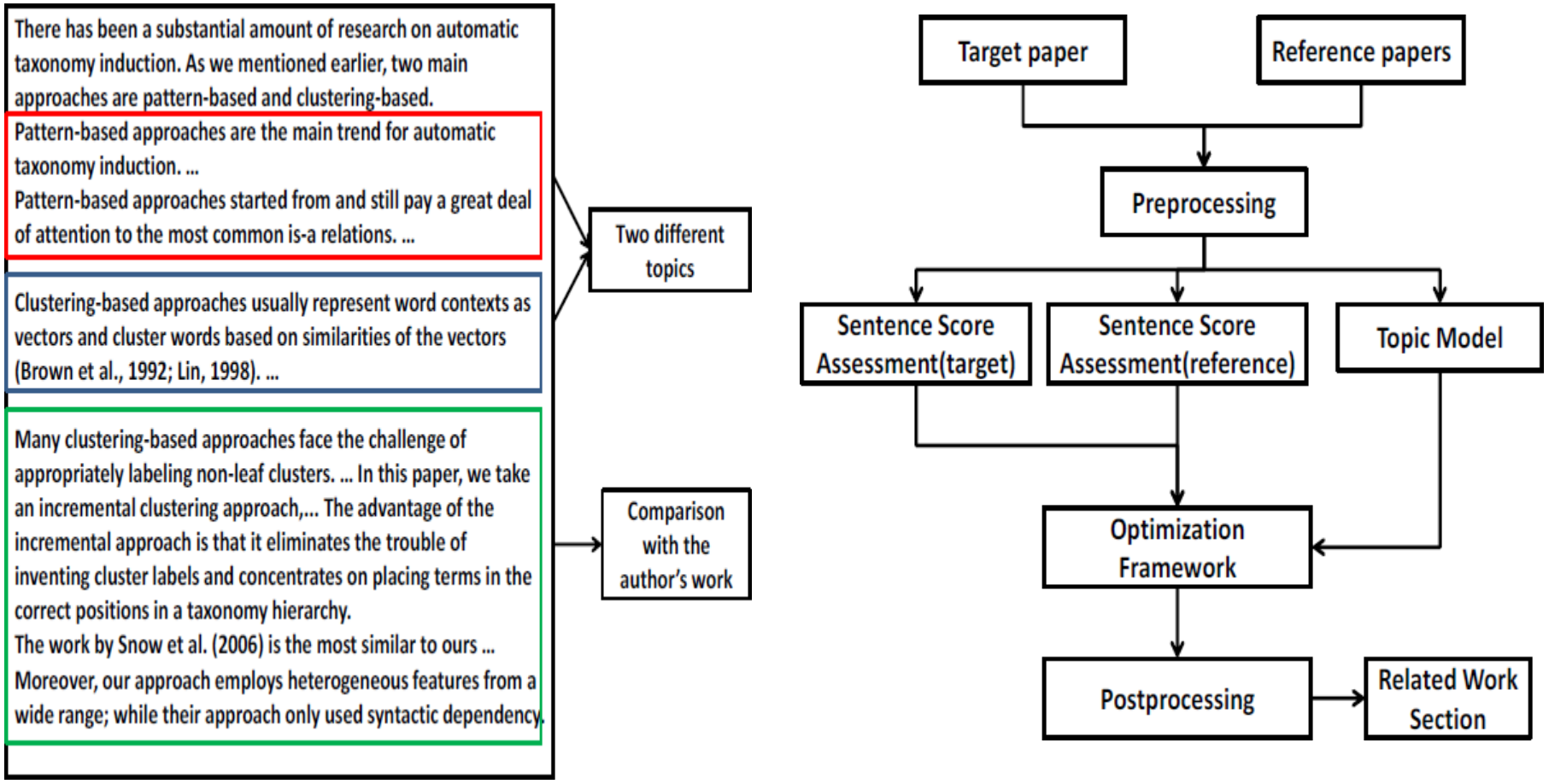
3 TRUSTED OBSERVERS AND THE MONITORING ARCHITECTURE (The Monitoring Architecture)

- Normative Systems
  - In this section we describe our architecture for monitoring the behaviours of agents deployed in normative systems.
  - Notice that there is nothing in the specification of the monitor that ties it to a particular normative system.

3 TRUSTED OBSERVERS AND THE MONITORING ARCHITECTURE (Motivating Trusted Observers)

- Enforcement Mechanisms
  - Enforcement mechanisms are required to motivate agent compliance with norms.
- Notification Message
  - Fulfilment of this obligation can be recognised by observing for F 's sending of a notification message to S, informing the latter that payment has been made.
  - Unlike the case of the notification message sent by F , no gain accrues to the bank if the bank mis-reports.

学术文献摘要之相关工作章节自动撰写



Hu, Yue, and Xiaojun Wan. "Automatic Generation of Related Work Sections in Scientific Papers: An OptimizationApproach." EMNLP2014.

跨语言摘要

- **基于短语的跨语言摘要**
  - **借鉴基于短语的机器翻译技术**
  - **每个翻译得到的句子的得分**

$$F(s) = \sum_{p \in s} d_0 g(p) + bg(s) + \eta\, dist(y(s))$$

g(p): 短语p的文档频率;
bg(s): 句子s的bigram值
Dist(y(s)): 句子s的distortion penalty
η <0
d0 : damping factor

  - **摘要的得分**

$$F(S) = \sum_{p \in S} \sum_{i=1}^{count(p,S)} d^{i-1} g(p) + \sum_{s \in S} bg(s) + \eta \sum_{s \in S} dist(y(s))$$

  - **利用贪心算法进行句子选择与摘要抽取**
  - **可在句子选择中进行句子压缩**

Jin-ge Yao, Xiaojun Wan and Jianguo Xiao. "Phrase-based Compressive Cross-Language Summarization." In *EMNLP 2015*.

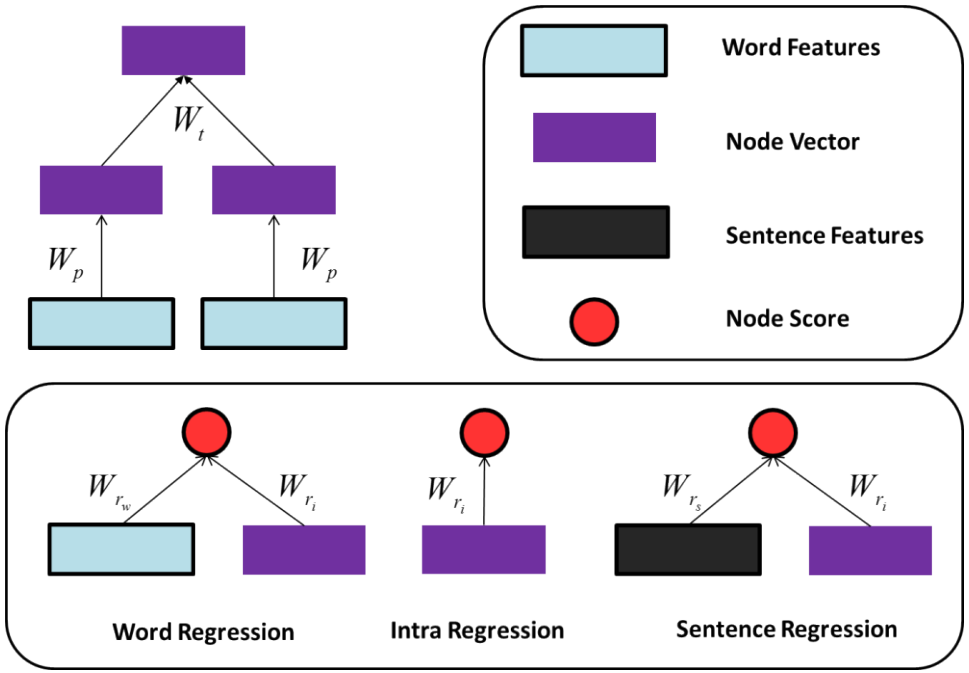
文本摘要与深度学习

- **相比其它NLP任务，深度学习技术较晚&较少应用于文本摘要任务**
  - **任务的特殊性：子集选择问题/压缩问题**
  - **数据规模（尤其是多文档摘要任务）**
  - **答案的相对发散与不一致性**
  - **长文档的语义编码**

- **目前已有多种尝试，但总体性能提升并不明显**

文本摘要与深度学习

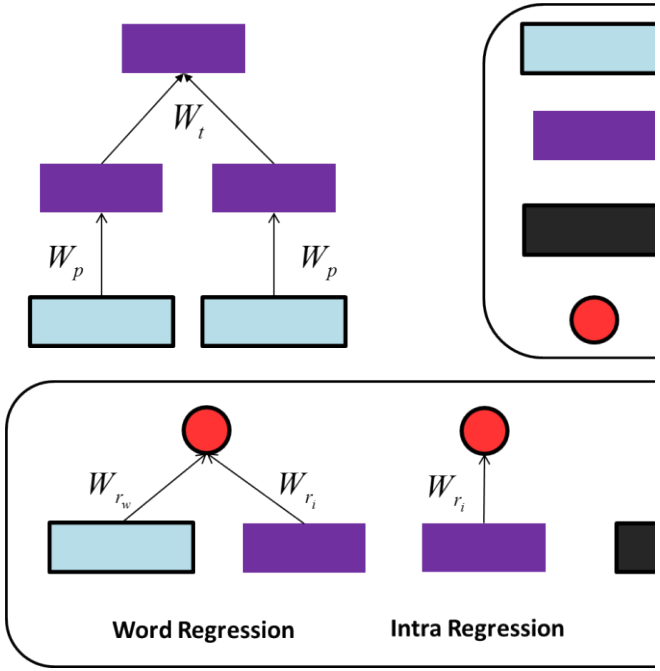
- **业界的主要尝试**
  - **在句子排序过程中使用RNN：R2N2**
    - **基于句子的句法树结构利用RNN自底向上获得每个节点的向量表示，同时结合传统特征进行重要性回归学习**



Cao, Ziqiang, et al. "Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization." AAAI. 2015.

文本摘要与深度学习

- **业界的主要尝试**
  - **在句子排序过程中使用RI**
    - **基于句子的句法树结构**
      示，同时结合传统特征



Year	System	ROUGE-1	ROUGE-2
2001	Peer T	33.06	**8.06**
	ClusterHITS*	**37.42**	6.81
	LexRank	33.22	5.76
	Ur	34.28	6.66
	Sr	34.06	6.65
	U+Sr	33.98	6.54
	R2N2_GA	35.88	7.64
	R2N2_ILP	36.91	7.87
2002	Peer 26	35.15	7.64
	ClusterCMRW*	**38.55**	8.65
	LexRank	35.09	7.51
	Ur	34.16	7.66
	Sr	34.23	7.81
	U+Sr	35.13	8.02
	R2N2_GA	36.84	8.52
	R2N2_ILP	37.96	**8.88**
2004	Peer 65	37.88	9.18
	REGSUM*	38.57	9.75
	LexRank	37.92	8.78
	Ur	37.22	9.15
	Sr	36.72	9.10
	U+Sr	37.62	9.31
	R2N2_GA	38.16	9.52
	R2N2_ILP	**38.78**	**9.86**

Table 3: Comparison results (%) on DUC datasets. "*" indicates the score is derived from the original paper.

Cao, Ziqiang, et al. "Ranking with Rec
Document Summarization." AAAI. 201

文本摘要与深度学习

- **业界的主要尝试**
  - **sequence-to-sequence模型/encoder-decoder框架及其变体**
    - **应用于句子压缩(句子摘要)任务**
      - **词序列=>选择标签[0,1]序列**
      - **词序列=>词序列**

Filippova, Katja, et al. "Sentence compression by deletion with lstms." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.

Rush, Alexander M., Sumit Chopra, and Jason Weston. "A neural attention model for abstractive sentence summarization." arXiv preprint arXiv:1509.00685 (2015).

Sumit Chopra, Michael Auli and Alexander M. Rush. "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks", NAACL2016.

文本摘要与深度学习

- **业界的主要尝试**
  - **sequence-to-sequence模型/encoder-decoder框架及其变体**
    - **应用于观点摘要任务**
      - **词序列=>词序列**

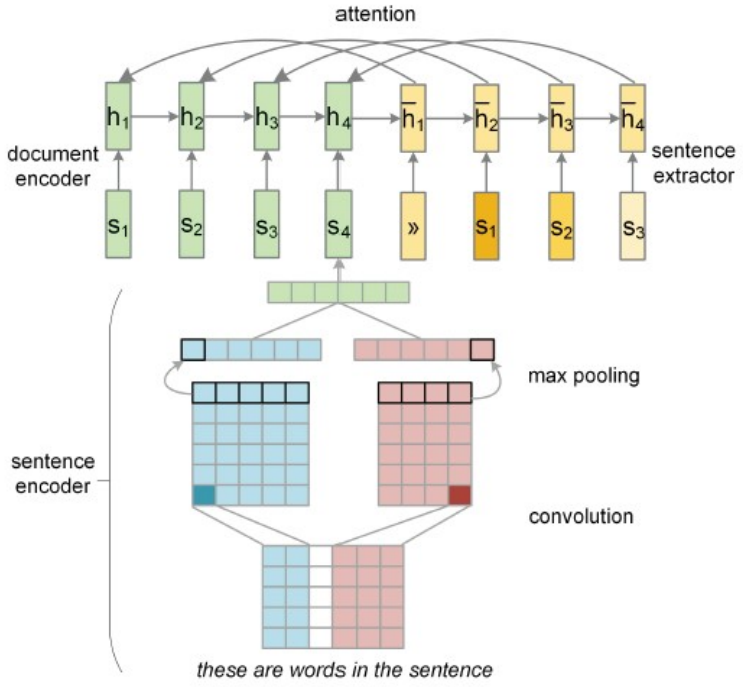
Wang, Lu, and Wang Ling. "Neural Network-Based Abstract Generation for Opinions and Arguments." NAACL 2016.

    - **应用于单文档摘要任务**
      - **句子序列=>句子选择标签[0,1]序列，并进一步预测词序列**

Cheng, Jianpeng, and Mirella Lapata. "Neural Summarization by Extracting Sentences and Words." arXiv preprint arXiv:1603.07252 (2016).

文本摘要与深度学习

- **业界的主要尝试**
  - **应用于文档摘要任务**
    - **句子序列=>句子选择标签[0,1]序列，并进一步预测词序列**
    - **基于DailyMail自动获取了数十万的训练语料**



	Models	ROUGE-1	ROUGE-2	ROUGE-L
DUC 2002	LEAD	43.6	21.0	40.2
	LREG	43.8	20.7	40.3
	ILP	45.4	21.3	42.8
	NN-ABS	15.8	5.2	13.8
	TGRAPH	48.1	**24.3**	—
	URANK	**48.5**	21.5	—
	NN-SE	47.4	23.0	**43.5**
	NN-WE	27.0	7.9	22.8

	Models	ROUGE-1	ROUGE-2	ROUGE-L
DailyMail	LEAD	36.4	15.3	29.4
	LREG	35.5	14.9	28.9
	NN-ABS	9.1	2.9	7.6
	NN-SE	**37.2**	**18.1**	**29.9**
	NN-WE	18.8	6.5	10.2

Table 1: ROUGE evaluation (%) on DUC-2002 and DailyMail corpora.

Cheng, Jianpeng, and Mirella Lapata. "Neural Summarization by Extracting Sentences and Words." arXiv preprint arXiv:1603.07252 (2016).

文本摘要技术的发展趋势

- 深度学习方法值得尝试，但短期内对文档摘要性能不会有 实质性的提高

- 对文本摘要的新需求将越来越多，研究将更加发散
  - **数字出版与移动互联网**

- 与人机交互技术相结合（尤其是对于移动阅读）
  - **Summly的成功：利用摘要与交互技术改变阅读方式**

- 与可视化技术相结合
  - **“一图胜千言”**

文本分类

Outline

- 文本分类的定义和应用
- 文本分类的方法
- 文本分类的评估指标
- 文本分类的其他方向
- 参考文献和资源

文本分类的定义和应用

定义

- 给定分类体系，将文本分到某个或者某几个类别中。
  - 分类体系一般人工构造
    - 政治、体育、军事
    - 中美关系、恐怖事件
  - 分类系统可以是层次结构，如yahoo!
  - 分类模式
    - 2类问题，属于或不属于(binary)
    - 多类问题，多个类别(multi-class)，可拆分成2类问题
    - 一个文本可以属于多类(multi-label)
  - 这里讲的分类主要基于内容
  - 很多分类体系: Reuters分类体系、中图分类

应用

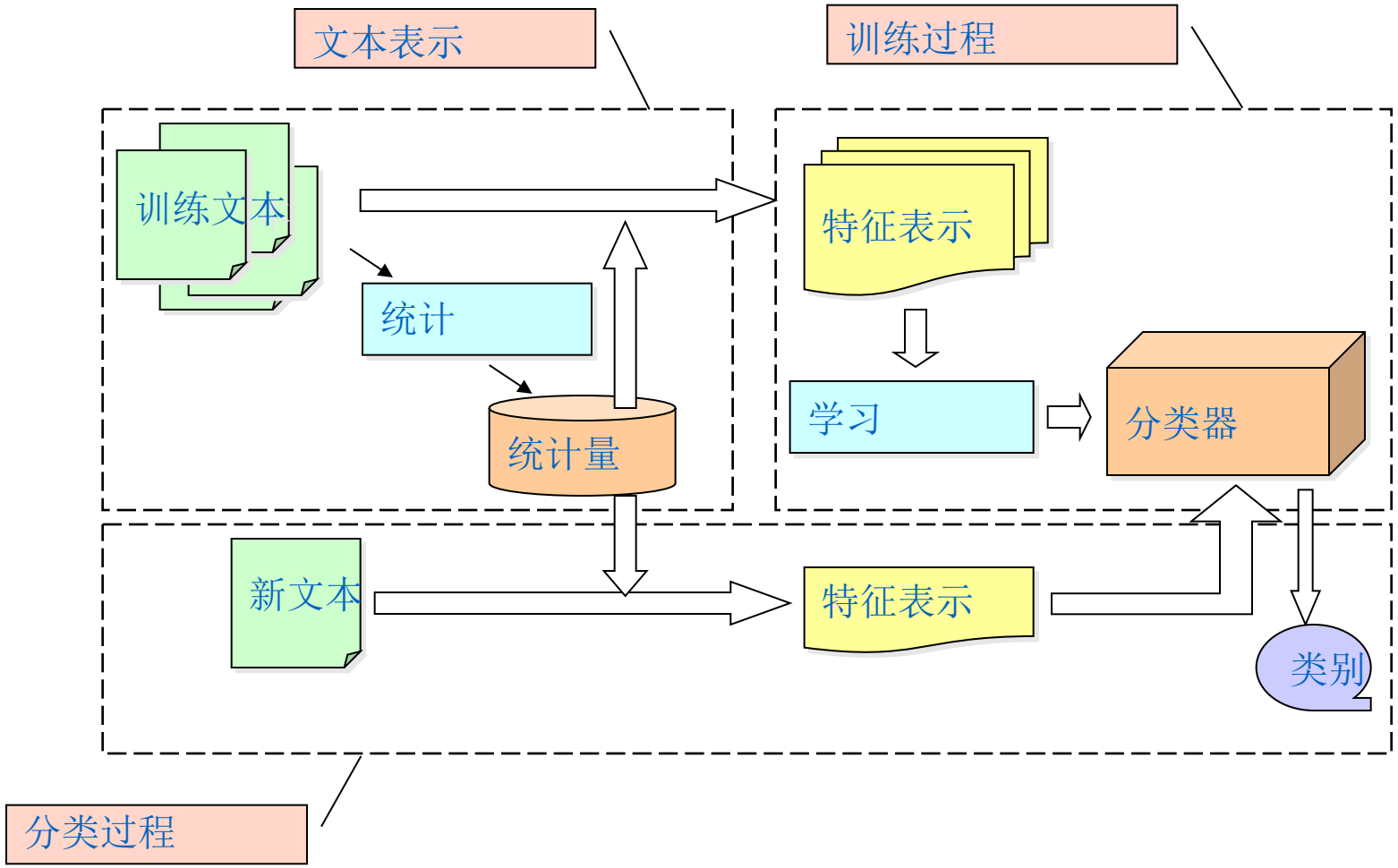
- 垃圾邮件的判定(spam or not spam)
  - 类别 {spam, not-spam}
- 新闻出版按照栏目分类
  - 类别 {政治,体育,军事,…}
- 词性标注
  - 类别 {名词,动词,形容词,…}
- 词义排歧
  - 类别 {词义1,词义2,…}
- 计算机论文的领域
  - 类别 ACM system
    - H: information systems
    - H.3: information retrieval and storage

文本分类的方法

人工方法和自动方法

- 人工方法
  - 结果容易理解
    - 足球 and 联赛→体育类
  - 费时费力
  - 难以保证一致性和准确性(40%左右的准确率)
  - 专家有时候凭空想象
  - 知识工程的方法建立专家系统(80年代末期)
- 自动的方法(学习)
  - 结果可能不易理解
  - 快速
  - 准确率相对高(准确率可达60%或者更高)
  - 来源于真实文本，可信度高

文本分类的过程



特征抽取

- 预处理
  - 去掉html一些tag标记
  - (英文)禁用词(stop words)去除、词根还原(stemming)
  - (中文)分词、词性标注、短语识别、…
  - 词频统计
    - $TF_{i,j}$: 特征i在文档j中出现次数，词频(Term Frequency)
    - DF_i:所有文档集合中出现特征i的文档数目，文档频率(Document Frequency)
  - 数据清洗：去掉不合适的噪声文档或文档内垃圾数据
- 文本表示
  - 向量空间模型(Vector Space Model)
- 降维技术
  - 特征选择(Feature Selection)
  - 特征重构(Re-parameterisation，如LSI、LDA)

文本表示

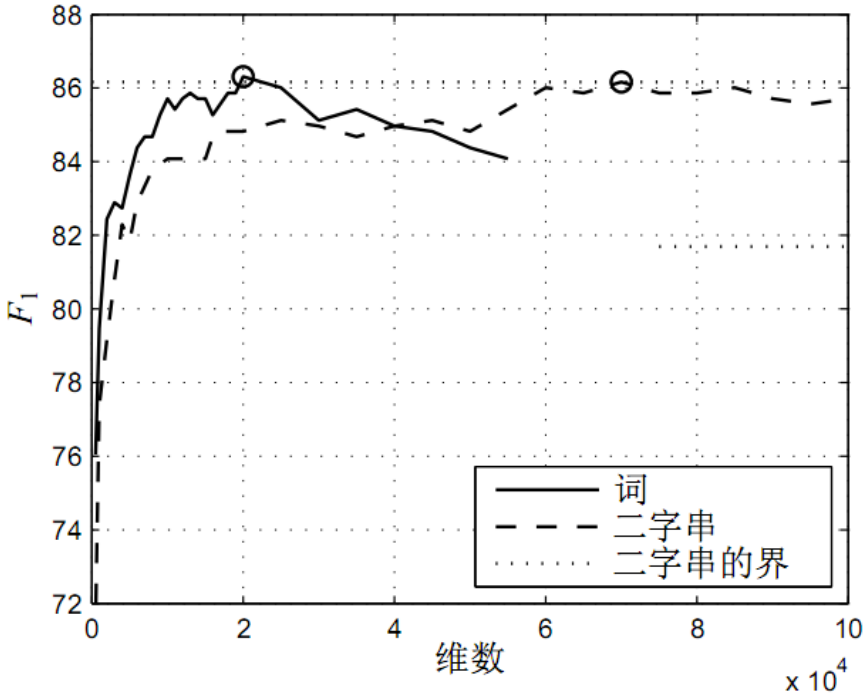
- 向量空间模型(Vector Space Model)
  - M个无序标引项t_i(特征)，词根/词/短语/其他
  - 假设所有特征独立
  - 每个文档d_j可以用标引项向量来表示
    - $(a_{1j},a_{2j},\ldots,a_{Mj})$
  - 权重计算，N个训练文档
    - $A_{M*N}= (a_{ij})$
  - 相似度比较
    - Cosine计算
    - 内积计算

Term的粒度

- Character，字：中
- Word，词：中国
- Phrase，短语：中国人民银行
- Concept，概念
  - 同义词：开心 高兴 兴奋
  - 相关词cluster，word cluster：鸟巢/水立方/奥运
- N-gram，N元组：中国 国人 人民 民银 银行
- 某种规律性模式：比如某个窗口中出现的固定模式
- 中文文本分类使用那种粒度？

Term粒度—中文

- 词特征 V.S. Bigram特征
  - 中文分词？更困难的学术问题
  - Bigram？简单粗暴
- 假设分词100%准确
  - 在低维度达到更好的结果
  - 现实中不可能的☹



注：左侧的圆圈标示词曲线的峰值，右侧的圆圈及上方的水平点线标示二字串曲线的峰值，下方的水平点线标示二字串在更高维时的性能下限。

Term粒度—中文

- ICTCLAS分词V.S. Bigram
  - 低维度：词 > Bigram
  - 高维度 ：Bigram > 词
  - 词的数目有限
  - Bigram特征数目更多，可以提供更多的特征
- So, 实用性角度：分词

  研究角度：Bigram

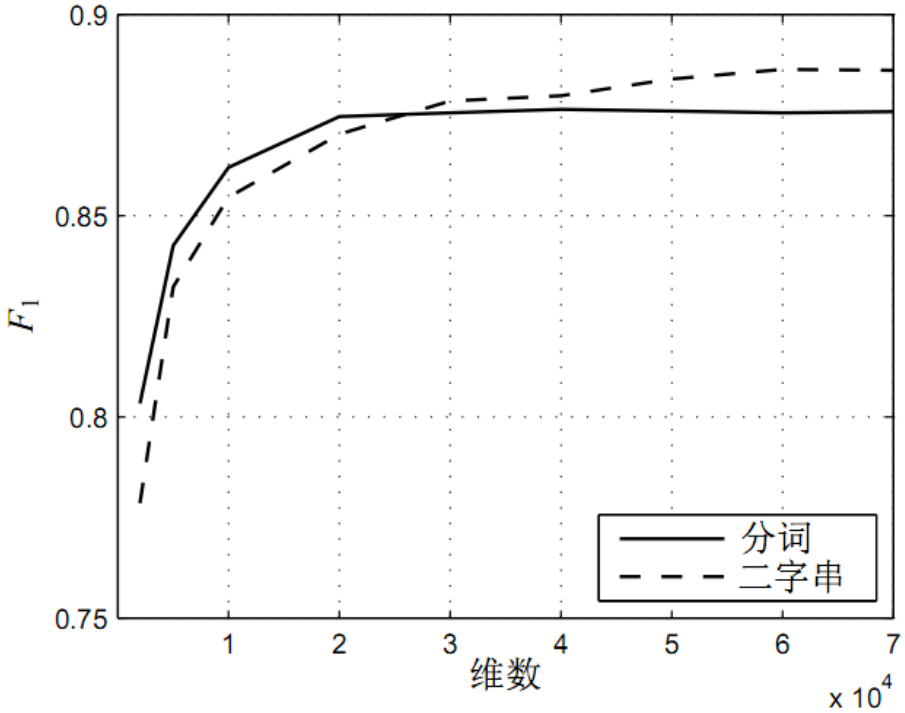


图 2.6　分词与二字串比较

权重计算方法

- 布尔权重(Boolean weighting)
  - a_{ij}=1(TF_{ij}>0) or (TF_{ij}=0)0
- TFIDF型权重
  - TF: $a_{ij}=TF_{ij}$
  - TF*IDF: $a_{ij}=TF_{ij}*\log(N/DF_i)$
  - TFC: 对上面进行归一化
  - LTC: 降低TF的作用
- 基于熵概念的权重(Entropy weighting)
  - 称为term i的某种熵
  - 如果term分布极度均匀：熵等于-1
  - 只在一个文档中出现：熵等于0

$$a_{ij}=\frac{TF_{ij}*\log(N/DF_i)}{\sqrt{\sum_k[TF_{kj}*\log(N/DF_k)]^2}}$$

$$a_{ij}=\frac{\log(TF_{ij}+1.0)*\log(N/DF_i)}{\sqrt{\sum_k[\log(TF_{kj}+1.0)*\log(N/DF_k)]^2}}$$

$$a_{ij}=\log(TF_{ij}+1.0)*\left(1+\frac{1}{\log N}\sum_{j=1}^{N}[\frac{TF_{ij}}{DF_i}\log(\frac{TF_{ij}}{DF_i})]\right)$$

特征选择(1)

- 基于DF
  - Term的DF小于某个阈值去掉(太少，没有代表性)
  - Term的DF大于某个阈值也去掉(太多，没有区分度)
- 信息增益(Information Gain, IG)：该term为整个分类所能提供的信息量(不考虑任何特征的熵和考虑该特征后的熵的差值)

$$\begin{aligned}\text{Gain(t)} &= \text{Entropy}(S) - \text{Expected Entropy}(S_t) \\ &= \{-\sum_{i=1}^{M} P(c_i)\log P(c_i)\} - \\ &\quad [P(t)\{-\sum_{i=1}^{M} P(c_i \mid t)\log P(c_i \mid t)\} + \\ &\quad P(\bar{t})\{-\sum_{i=1}^{M} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t})\}]\end{aligned}$$

特征选择(2)

- term的某种熵：该值越大，说明分布越均匀，越有可能出现在较多的类别中(区分度差)；该值越小，说明分布越倾斜，词可能出现在较少的类别中(区分度好)

$$Entropy(t) = -\sum_i P(c_i \mid t) \log P(c_i \mid t)$$

- 相对熵(not 交叉熵)：也称为KL距离(Kullback-Leibler divergence)，反映了文本类别的概率分布和在出现了某个特定词汇条件下的文本类别的概率分布之间的距离，该值越大，词对文本类别分布的影响也大。

$$CE(t) = \sum_i P(c_i \mid t) \log \frac{P(c_i \mid t)}{P(c_i)}$$

特征选择(3)

- χ^2 统计量：度量两者(term和类别)独立性的缺乏程度，χ^2 越大，独立性越小，相关性越大(若AD<BC,则类和词独立, N=A+B+C+D)

	c	~c
t	A	B
~t	C	D

$$\chi^2(t,c)=\frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

$$\chi^2{}_{AVG}(t)=\sum_{i=1}^{m}P(c_i)\chi^2(t,c_i)$$

$$\chi^2{}_{MAX}(t)=\max_{i=1}^{m}\{\chi^2(t,c_i)\}$$

- 互信息(Mutual Information)：MI越大t和c共现程度越大

$$I(t,c)=\log\frac{P(t\wedge c)}{P(t)P(c)}=\log\frac{P(t\mid c)}{P(t)}=\log\frac{A\times N}{(A+C)(A+B)}$$

$$I_{AVG}(t)=\sum_{i=1}^{m}P(c_i)I(t,c_i)\qquad I_{MAX}(t)=\max_{i=1}^{m}P(c_i)I(t,c_i)$$

特征重构

- 隐性语义索引(Latent Semantic Index)
  - 奇异值分解(SVD)：$A=(a_{ij})=U\Sigma V^T$
    - A_{M*N}, U_{M*R}, Σ_{R*R}(对角阵), V_{N*R}, R<=MIN(M,N)
  - 取Σ对角上的前k个元素，得Σ_k
    - $A_k= U_k\Sigma_k V_k^T$, U_k由U的前k列组成，V_k由V的前k列组成
    - 文档d在LSI对应的向量$d'=d^TU_k\Sigma^{-1}$
- Latent Dirichlet allocation(LDA)
  - Topic Model
  - Bag-Of-Words表示 -> Topic表示

自动文本分类方法

- Rocchio方法
- Naïve Bayes
- kNN方法
- 决策树方法decision tree
- Decision Rule Classifier
- The Widrow-Hoff Classifier
- 神经网络方法Neural Networks
- 支持向量机SVM
- 基于投票的方法(voting method)

Rocchio方法

- 可以认为类中心向量法是它的特例
  - Rocchio公式

$$w'_{jc} = \alpha w_{jc} + \beta \frac{\sum_{i \in C} x_{ij}}{n_C} - \gamma \frac{\sum_{i \notin C} x_{ij}}{n - n_C}$$

w'_{jc}：类C中心向量的权重；n_C：训练样本中正例个数；x_{ij}：文档向量的权重

  - 分类

$$CSV_c(d_i) = \mathrm{w_c} \cdot \mathrm{x}_i = \frac{\sum w_{cj} \cdot x_{ij}}{\sqrt{\sum w_{cj}^{\ 2}} \sqrt{\sum x_{ij}^{\ 2}}}$$

Naïve Bayes

Bayes公式

$$P(c_j \mid d_i) = \frac{P(d_i \mid c_j)P(c_j)}{P(d_i)} \propto P(d_i \mid c_j)P(c_j)$$

$$P(d_i \mid c_j) = \prod_{k=1}^{r} P(w_{ik} \mid c_j), \quad \text{独立性假设}$$

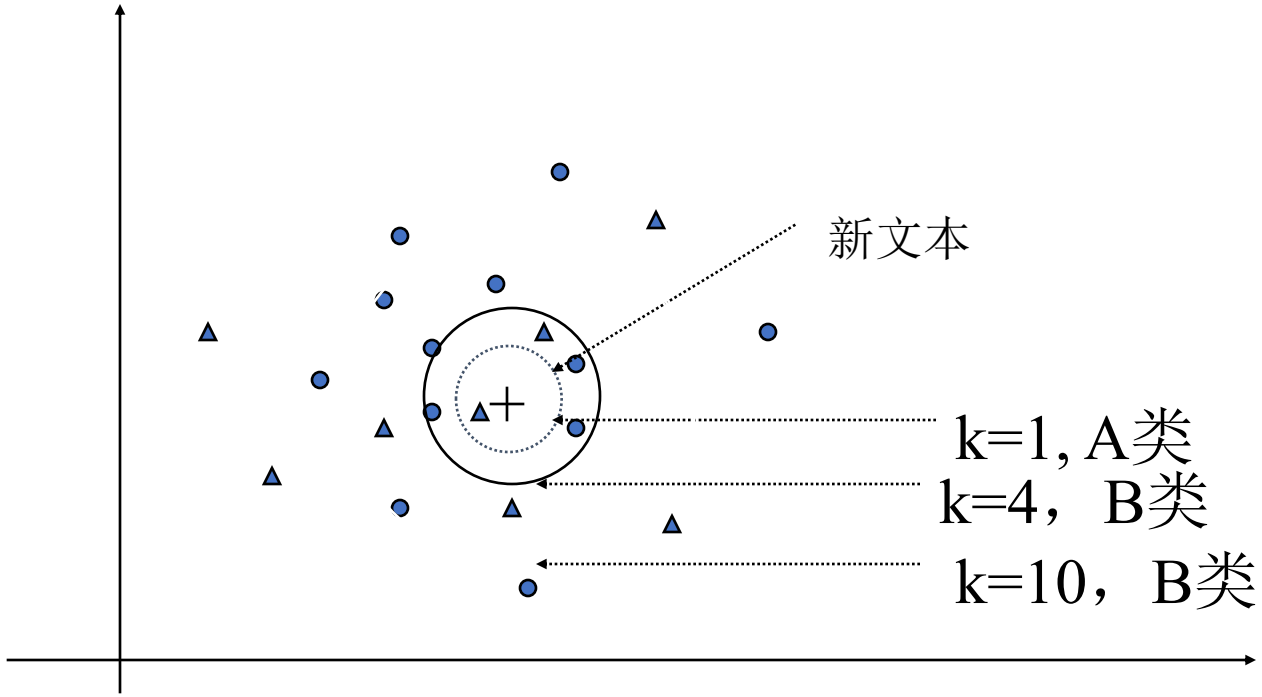
参数计算

$$P(c_j) = \frac{c_j\text{的文档个数}}{\text{总文档个数}} = \frac{N(c_j)}{\sum_k N(c_k)} \approx \frac{1 + N(c_j)}{|c| + \sum_{k=1} N(c_k)}$$

$$P(w_i \mid c_j) = \frac{w_i\text{在}c_j\text{类别文档中出现的次数}}{\text{在}c_j\text{类所有文档中出现的词的次数}} \approx \frac{1 + N_{ij}}{\text{不同词个数} + \sum_k N_{kj}}$$

kNN方法

- 一种Lazy Learning, Example-based Learning



带权重计算，计算权重和最大的类。k常取3或者5。

决策树方法

- 构造决策树
  - CART
  - C4.5 (由ID3发展而来)
  - CHAID
- 决策树的剪枝(pruning)

Decision Rule Learning

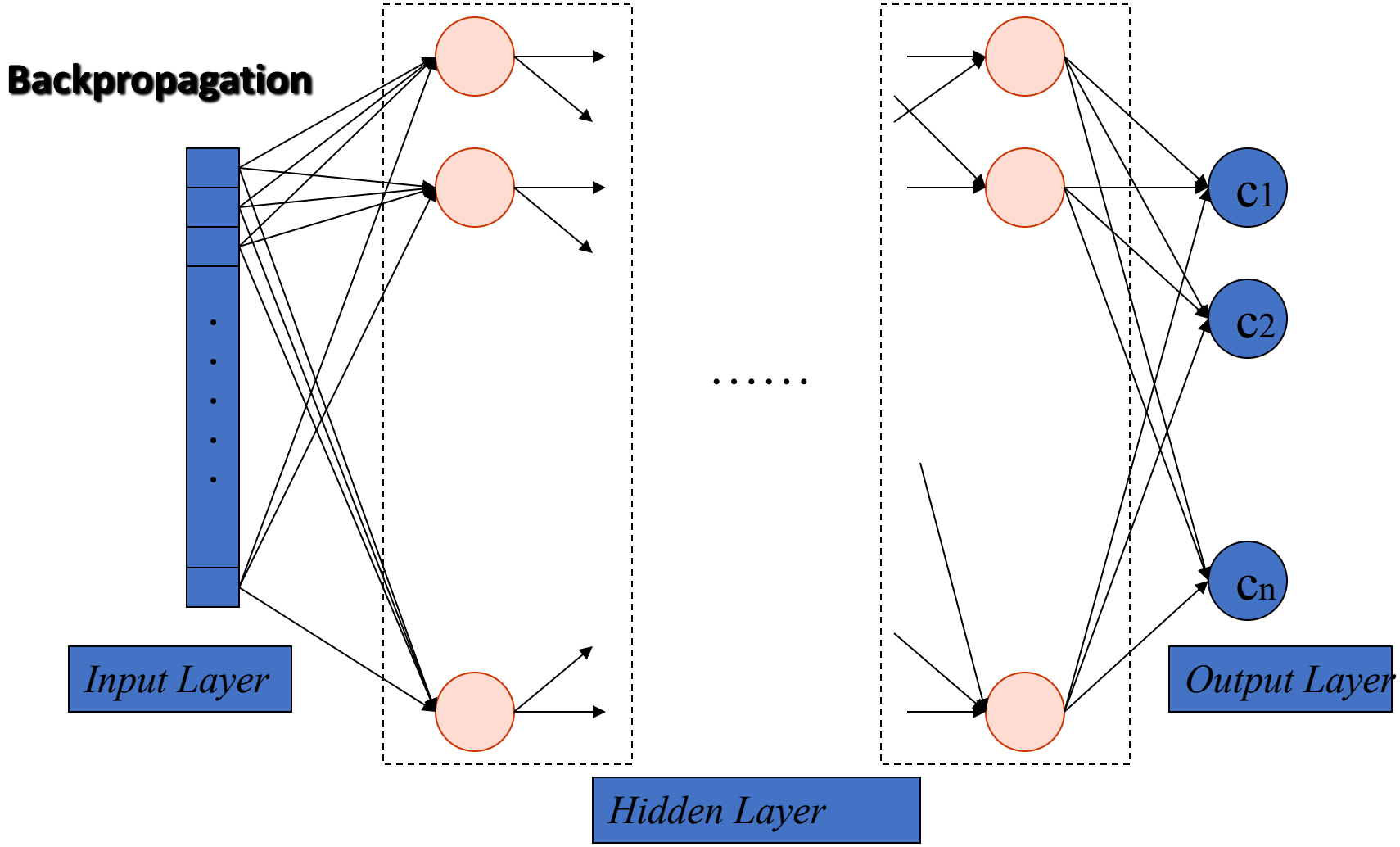
学习到如下规则

wheat & form → WHEAT
wheat & commodity → WHEAT
bushels & export → WHEAT
wheat & agriculture → WHEAT
wheat & tonnes → WHEAT
wheat & winter & ~soft → WHEAT

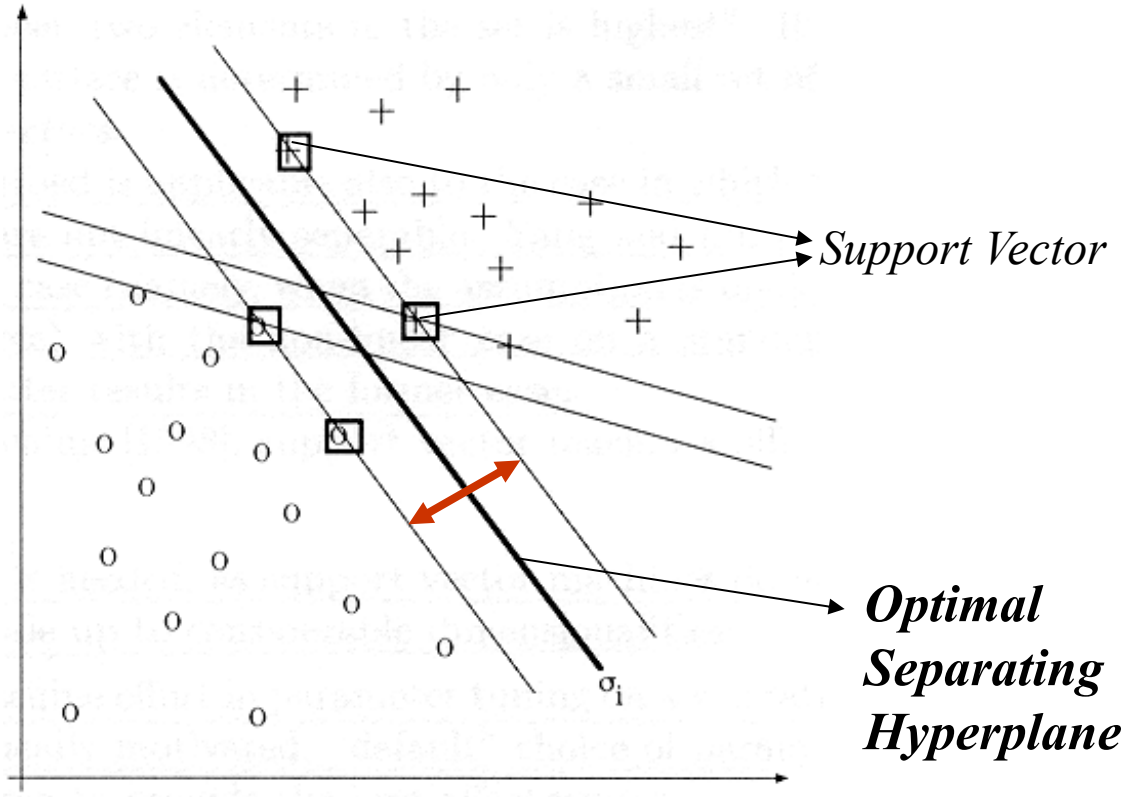
(粗糙集)RoughSet

逻辑表达式(AQ11算法)

Neural Network



支持向量机
Support Vector Machine



基于投票的方法

- Bagging方法
  - 训练R个分类器f_i，分类器之间其他相同就是参数不同。其中f_i是通过从训练集合中(N篇文档)随机取(取后放回)N次文档构成的训练集合训练得到的。
  - 对于新文档d，用这R个分类器去分类，得到的最多的那个类别作为d的最终类别
- Boosting方法
  - 类似Bagging方法，但是训练是串行进行的，第k个分类器训练时关注对前k-1分类器中错分的文档，即不是随机取，而是加大取这些文档的概率(加大对错分样本的学习能力)
  - AdaBoost

文本分类的评估指标

分类方法的评估

- 邻接表

	真正对的	错误
标YES	a	b
标NO	c	d

- 每个类
  - Precision=a/(a+b), Recall=a/(a+c), fallout=b/(b+d)=false alarm rate, accuracy=(a+d)/(a+b+c+d), error=(b+c)/(a+b+c+d)=1-accuracy, miss rate=1-recall
  - $F=(\beta^2+1)p.r/(\beta^2p+r)$
  - Break Even Point, BEP, p=r的点
  - 如果多类排序输出，采用interpolated 11 point average precision
- 所有类：
  - 宏平均:对每个类求值，然后平均
  - 微平均:将所有文档一块儿计算，求值

其他分类方法

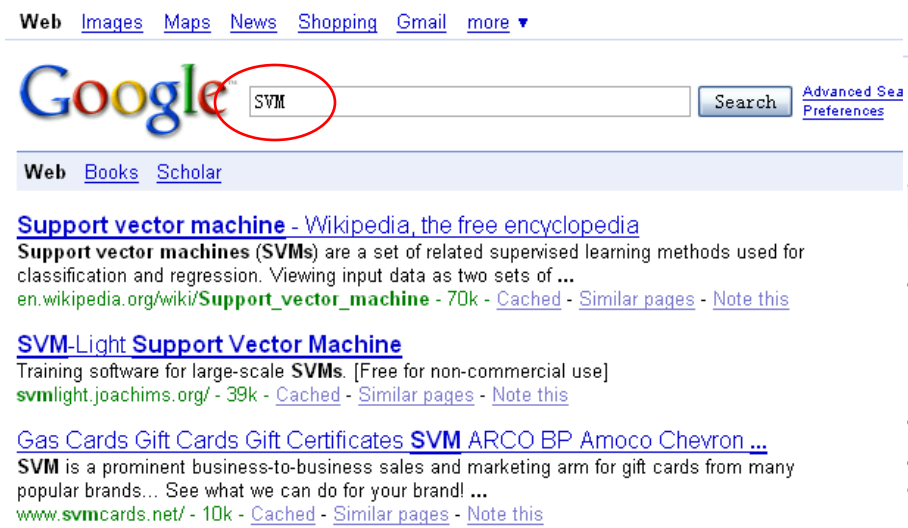
- Regression based on Least Squares Fit (1991)
- Nearest Neighbor Classification (1992) *
- Bayesian Probabilistic Models (1992) *
- Symbolic Rule Induction (1994)
- Decision Tree (1994) *
- Neural Networks (1995)
- Rocchio approach (traditional IR, 1996) *
- Support Vector Machines (1997)
- Boosting or Bagging (1997)*
- Hierarchical Language Modeling (1998)
- First-Order-Logic Rule Induction (1999)
- Maximum Entropy (1999)
- Hidden Markov Models (1999)
- Error-Correcting Output Coding (1999)
- …

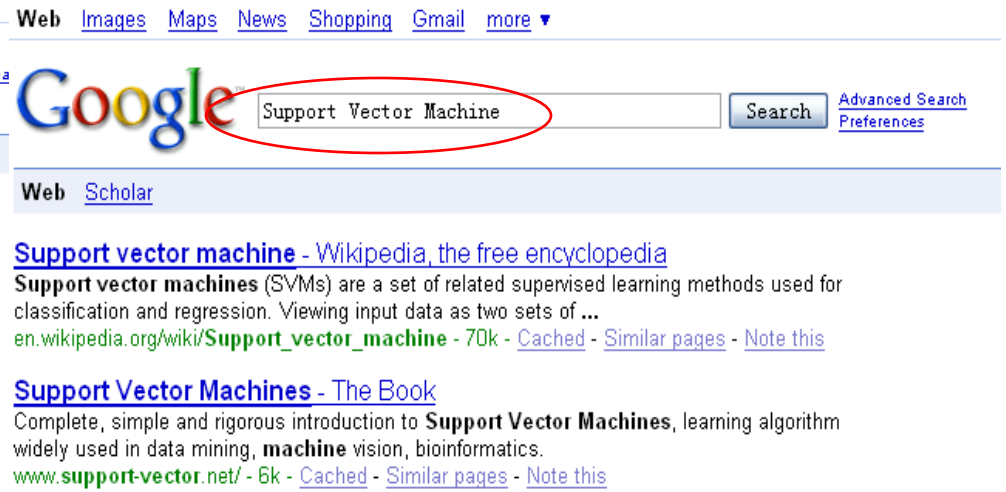
传统文本分类研究方向

- 特征选择
- 权重计算
- 不平衡数据集分类
- 训练集样本很少(半监督学习)
- Active-Learning：加入人工的因素
- 基本上文本分类作为检验新的机器学习方法的平台

新方向

- 短文本分类
  - 最大的问题：信息缺失
  - Ask Google Snippet





开方测试问题

- 论文中的指标都是在封闭训练测试上计算
- Web上的文本错综复杂，不可能有统一的分类体系
- 在训练集合A上的模型，自适应的转移到集合B中的文本分布？
- Transfer Learning
- 主要问题在于成本较高

其他一些问题

- 多类别数目分类问题：
  - 比如类别数有成百上千的情况
  - SVM？训练时一般采用One V.S. One方法
  - 如果一定要选，Naïve Bayes方法更鲁棒
- 分类速度：
  - 实用的角度不可能采用paper中的方法
  - 一般在速度和效果中寻求Tradeoff

参考文献

文献及其他资源

- Papers
  - K. Aas and L. Eikvil. *Text categorisation: A survey*. Technical report, Norwegian Computing Center, June 1999
  - Yiming Yang and Xin Liu. 1999. “A re-examination of text categorization methods.” *SIGIR 1999*
  - *A Survey on Text Categorization*, NLP Lab, Korean U.
  - 黄萱菁等，*独立于语种的文本分类方法*，中文信息学报，2000年第6期
- Software:
  - Rainbow http://www-2.cs.cmu.edu/~mccallum/bow/
  - BoosTexter http://www.research.att.com/~schapire/BoosTexter/
  - TiMBL http://ilk.kub.nl/software.html#timbl
  - C4.5 http://www.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html
- Corpus
  - http://www.cs.cmu.edu/~textlearning

文本聚类

一、概述

二、相似性度量

三、划分方法

四、层次聚类方法

五、基于密度的聚类

六、基于网格方法

七、基于模型方法

八、蚁群聚类方法

十、粒度计算

十一、实例分析与计算机实现

概 述

- 无监督学习不要求对数据进行事先标定，在数据的分类结构未知时，按照事物的某些属性，把事物聚集成类，使类间的相似性尽量小，类内相似性尽量大。利用无监督学习期望能够发现数据集中自身隐藏的内蕴结构信息。
- 无监督学习也称聚类分析。 无监督学习源于许多研究领域，受到很多应用需求的推动。例如，
- 在复杂网络分析中，人们希望发现具有内在紧密联系的社团
- 在图像分析中，人们希望将图像分割成具有类似性质的区域
- 在文本处理中，人们希望发现具有相同主题的文本子集
- 在有损编码技术中，人们希望找到信息损失最小的编码
- 在顾客行为分析中，人们希望发现消费方式类似的顾客群，以便制订有针对性的客户管理方式和提高营销效率。这些情况都可以在适当的条件下归为聚类分析。

概 述

- “物以类聚，人以群分”。
- 一般的聚类算法是先选择若干个模式点作为聚类的中心。每一中心代表一个类别，按照某种相似性度量方法（如最小距离方法）将各模式归于各聚类中心所代表的类别，形成初始分类。然后由聚类准则判断初始分类是否合理，如果不合理就修改分类，如此反复迭代运算，直到合理为止。与监督学习不同，无监督法是边学习边分类，通过学习找到相同的类别，然后将该类与其它类区分开。

聚类分析

- **聚类分析(clustering analysis)**是将样品个体或指标变量按其具有的特性进行分类的一种统计分析方法。
  - 对样品进行聚类，称为样品(Q型)聚类分析。其目的是将分类不明确的样品按性质相似程度分成若干组，从而发现同类样品的共性和不同类样品间的差异。
  - 对指标进行聚类，称为指标（R型）聚类分析。其目的是将分类不明确的指标按性质相似程度分成若干组，从而在尽量不损失信息的条件下，用一组少量的指标来代替原来的多个指标（主成分分析？因子分析？）

聚类分析

典型的数据聚类基本步骤如下：

- (1)对数据集进行表示和预处理，包括数据清洗、特征选择或特征抽取；
- (2)给定数据之间的相似度或相异度及其定义方法；
- (3)根据相似度，对数据进行划分，即聚类；
- (4)对聚类结果进行评估。

相似性度量

如何刻画样品/（指标）变量间的亲疏关系或相似程度？

样品相似性的度量

变量相似性的度量

相似系数度量

相似系数体现对象间的相似程度，反映样本之间相对于某些属性的相似程度。确定相似系数有很多方法，这里列出一些常用的方法，可以根据实际问题选择使用。

设 为被分类对象的全体，以 表示每一对象 的特征数据。令$x_i, x_j \in O$, r_{ij}是x_i和 x_j之间的相似系数，满足以下条件：

- $r_{ij}=1 \Leftrightarrow x_i = x_j$
- $\forall x_i, x_j, r_{ij} \in [0,1]$
- $\forall x_i, x_j, r_{ij} = r_{ji}$

相似系数度量

1. 数量积法

$$r_{ij}=\begin{cases}1 & i=j;\\ \dfrac{1}{M}\sum_{k=1}^{m}x_{ik}x_{jk} & i\neq j.\end{cases}$$

其中，M为正数，满足 $M\geq\max_{i\neq j}(\sum_{k=1}^{m}x_{ik}x_{jk})$

相似系数度量

2、夹角余弦

两变量X_i与X_j看作p维空间的两个向量，这两个向量间的夹角余弦可用下式进行计算

$$\cos\theta_{ij}=\frac{\sum_{k=1}^{p}X_{ik}X_{jk}}{\sqrt{(\sum_{k=1}^{p}X_{ik}^2)(\sum_{k=1}^{p}X_{jk}^2)}}$$

显然，$|\cos\theta_{ij}|\leq 1$。

相似系数度量

3．相关系数

相关系数经常用来度量变量间的相似性。变量x_i与x_j的相关系数定义为

$$r_{ij}=\frac{\sum_{k=1}^{p}(X_{ik}-\bar{X}_i)(X_{jk}-\bar{X}_j)}{\sqrt{\sum_{k=1}^{p}(X_{ik}-\bar{X}_i)^2\sum_{k=1}^{p}(X_{jk}-\bar{X}_j)^2}}$$

显然也有，$|r_{ij}|\leq 1$。

相似系数度量

4．最大最小法

$$r_{ij}=\frac{\sum_{k=1}^{m}(x_{ik}\wedge x_{jk})}{\sum_{k=1}^{m}(x_{ik}\vee x_{jk})}$$

5．算术平均最小法

$$r_{ij}=\frac{2\sum_{k=1}^{m}(x_{ik}\wedge x_{jk})}{\sum_{k=1}^{m}(x_{ik}+x_{jk})}$$

相似系数度量

6．几何平均最小法

$$r_{ij}=\frac{\sum_{k=1}^{m}(x_{ik}\wedge x_{jk})}{\sum_{k=1}^{m}\sqrt{x_{ik}x_{jk}}}$$

7．绝对值指数法

$$r_{ij}=e^{-\sum_{k=1}^{m}|x_{ik}-x_{jk}|}$$

相似系数度量

8．指数相似系数法

$$r_{ij}=\frac{1}{m}\sum_{k=1}^{m}e^{-(x_{ik}-x_{jk})^2\,/\,s_k^2}$$

9．绝对值倒数法

$$r_{ij}=\begin{cases}1 & i=j\\ \dfrac{M}{\sum_{k=1}^{m}|x_{ik}-x_{jk}|} & i\neq j\end{cases}$$

划分方法

划分聚类方法(partitioning method，PAM)是给定一个有n个对象或元组的的数据库构建k个划分的方法。每个划分为一个类（或簇），并且$k \leq n$。每个类至少包含一个对象，每个对象必须属于而且只能属于一个类(模糊划分计算除外)。所形成的聚类将使得一个客观划分标准最优化，从而使得一个聚类中对象是“相似”的，而不同聚类中的对象是“不相似”的

K-means聚类分析

K-means法是麦奎因（MacQueen，1967）提出的，这种算法的基本思想是将每一个样品分配给最近中心（均值）的类中，具体的算法至少包括以下三个步骤：

(1)从*n*个数据对象随机选取*k*个对象作为初始簇中心。

(2)计算每个簇的平均值，并用该平均值代表相应的簇。

(3)计算每个对象与这些中心对象的距离，并根据最小距离重新对相应对象进行划分。

(4)转步骤(2)，重新计算每个(自变化)簇的平均值。这个过程不断重复直到某个准则函数不再明显变化或者聚类的对象不再变化为止。

K-means聚类分析

- 【例】假定我们对A、B、C、D四个样品分别测量两个变量和得到结果见表。

样品	变量	
	X_1	X_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

样品测量结果

试将以上的样品聚成两类。

K-means聚类分析

第一步：按要求取K=2，为了实施均值法聚类，我们将这些样品随意分成两类，比如（A、B）和（C、D），然后计算这两个聚类的中心坐标，见下表所示。

聚类	中心坐标	
	$\bar{X}_1$	$\bar{X}_2$
（A、B）	2	2
（C、D）	-1	-2

中心坐标是通过原始数据计算得来的，比如（A、 B）类的

$$\bar{X}_1 = \frac{5+(-1)}{2} = 2$$

K均值聚类分析

第二步：计算某个样品到各类中心的欧氏平方距离，然后将该样品分配给最近的一类。对于样品有变动的类，重新计算它们的中心坐标，为下一步聚类做准备。先计算A到两个类的平方距离：

$$d^2(A,(AB)) = (5-2)^2 + (3-2)^2 = 10$$

$$d^2(A,(CD)) = (5+1)^2 + (3+2)^2 = 61$$

由于A到（A、B）的距离小于到（C、D）的距离，因此A不用重新分配。计算B到两类的平方距离：

$$d^2(B,(AB)) = (-1-2)^2 + (1-2)^2 = 10$$

$$d^2(B,(CD)) = (-1+1)^2 + (1+2)^2 = 9$$

K均值聚类分析

- 由于B到（A、B）的距离大于到（C、D）的距离，因此B要分配给（C、D）类，得到新的聚类是（A）和（B、C、D）。更新中心坐标如下表所示。

聚类	中心坐标	
	$\overline{X}_1$	$\overline{X}_2$
（A）	5	3
（B、C、D）	-1	-1

更新后的中心坐标

K均值聚类分析

第三步：再次检查每个样品，以决定是否需要重新分类。计算各样品到各中心的距离平方，结果见下表。

聚类	样品到中心的距离平方			
	A	B	C	D
（A）	0	40	41	89
（B、C、D）	52	4	5	5

- 到现在为止，每个样品都已经分配给距离中心最近的类，因此聚类过程到此结束。最终得到K=2的聚类结果是A独自成一类，B、C、D聚成一类。

距离选择的原则

一般说来，同一批数据采用不同的距离公式，会得到不同的分类结果。产生不同结果的原因，主要是由于不同的距离公式的侧重点和实际意义都有不同。因此我们在进行聚类分析时，应注意距离公式的选择。通常选择距离公式应注意遵循以下的基本原则：

- （1）要考虑所选择的距离公式在实际应用中有明确的意义。如欧氏距离就有非常明确的空间距离概念。马氏距离有消除量纲影响的作用。
- （2）要综合考虑对样本观测数据的预处理和将要采用的聚类分析方法。如在进行聚类分析之前已经对变量作了标准化处理，则通常就可采用欧氏距离。
- （3）要考虑研究对象的特点和计算量的大小。样品间距离公式的选择是一个比较复杂且带有一定主观性的问题，我们应根据研究对象的特点不同做出具体分折。实际中，聚类分析前不妨试探性地多选择几个距离公式分别进行聚类，然后对聚类分析的结果进行对比分析，以确定最合适的距离测度方法。

层次聚类方法
(hierarchical method)

- **定义：对给定的数据进行层次的分解：**
- **分类：**

➢**凝聚方法（agglomerative）（自底向上）**
思想：一开始将每个对象作为单独的一组，然后根据同类相近，异类相异的原则，合并对象，直到所有的组合并成一个，或达到一个终止条件为止。

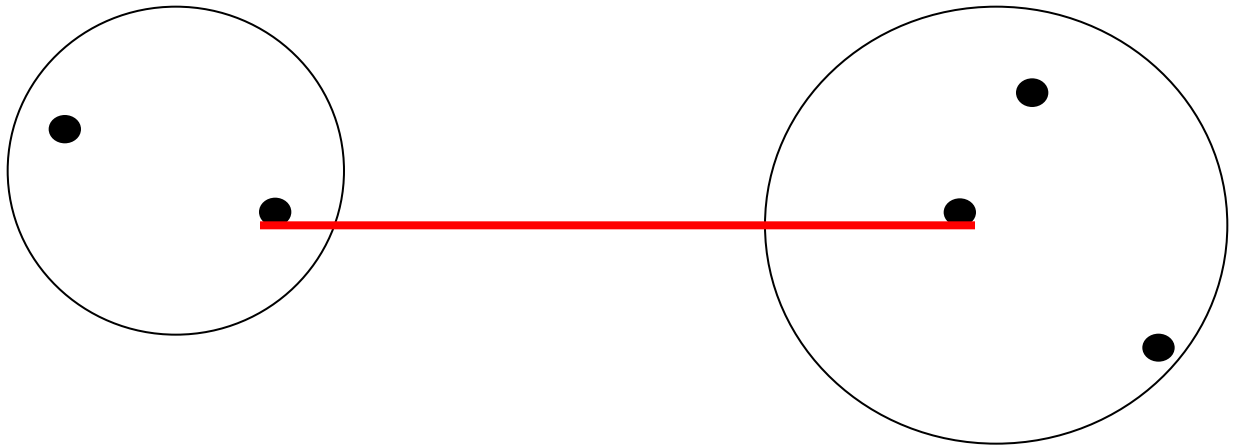
➢**分裂方法（divisive）（自顶向下）**
思想：一开始将所有的对象置于一类，在迭代的每一步中，一个类不断地分为更小的类，直到每个对象在单独的一个类中，或达到一个终止条件。

层次聚类方法
(hierarchical method)

- **特点：**
  - **类的个数不需事先定好**
  - **需确定距离矩阵**
  - **运算量要大，适用于处理小样本数据**

层次聚类方法

- 广泛采用的类间距离：
- 最小距离法（single linkage method）
  - 极小异常值在实际中不多出现，避免极大值的影响



最短距离法

1. 最短距离法

定义类与之间的距离为两类最近样品的距离，即为

$$D_{ij} = \min_{X_i \in G_i, X_j \in G_j} d_{ij}$$

设类与合并成一个新类记为，则任一类与的距离为

$$D_{kr} = \min_{X_i \in G_k, X_j \in G_r} d_{ij}$$

$$= \min\{ \min_{X_i \in G_k, X_j \in G_p} d_{ij}, \min_{x_i \in G_k, x_j \in G_q} d_{ij} \}$$

$$= \min\{D_{kp}, D_{kq}\}$$

最短距离法

- 最短距离法进行聚类分析的步骤如下：

（1）定义样品之间距离，计算样品的两两距离，得一距离阵记为$D_{(0)}$，开始每个样品自成一类，显然这时$D_{ij} = d_{ij}$。

（2）找出距离最小元素，设为D_{pq}，则将G_p和G_q合并成一个新类，记为G_r，即$G_r = \{G_p, G_q\}$。

（3）按（5.12）计算新类与其它类的距离。

（4）重复（2）、（3）两步，直到所有元素。并成一类为止。如果某一步距离最小的元素不止一个，则对应这些最小元素的类可以同时合并。

最短距离法

- 【例】设有六个样品，每个只测量一个指标，分别是1，2，5，7，9，10，试用最短距离法将它们分类。

（1）样品采用绝对值距离，计算样品间的距离阵$D_{(0)}$，见表

	G_1	G_2	G_3	G_4	G_5	G_6
G_1	0					
G_2	1	0				
G_3	4	3	0			
G_4	6	5	2	0		
G_5	8	7	4	2	0	
G_6	9	8	5	3	1	0

表

最短距离法

（2）$D_{(0)}$ 中最小的元素是$D_{12} = D_{56} = 1$，于是将G_1和G_2合并成G_7，G_5和G_6合并成G_8，并利用式计算新类与其它类的距离$D_{(1)}$，见下表：

	G_7	G_3	G_4	G_8
G_7	0			
G_3	3	0		
G_4	5	2	0	
G_8	7	4	2	0

表

最短距离法

（3）在$D_{(1)}$中最小值是$D_{34}=D_{48}=2$，由于G_4与G_3合并，又与G_8合并，因此G_3、G_4、G_8合并成一个新类G_9，其与其它类的距离$D_{(2)}$，见下表：

	G_7	G_9
G_7	0	
G_9	3	0

表

最短距离法

（4） 最后将G_7和G_9合并成G_{10}，这时所有的六个样品聚为一类，其过程终止。

上述聚类的可视化过程见下图所示，横坐标的刻度表示并类的距离。这里我们应该注意，聚类的个数要以实际情况所定，其详细内容将在后面讨论。

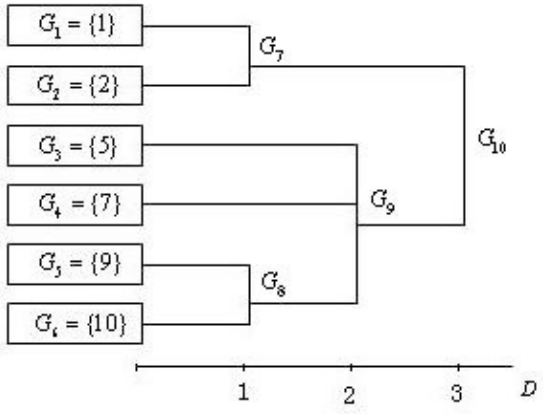
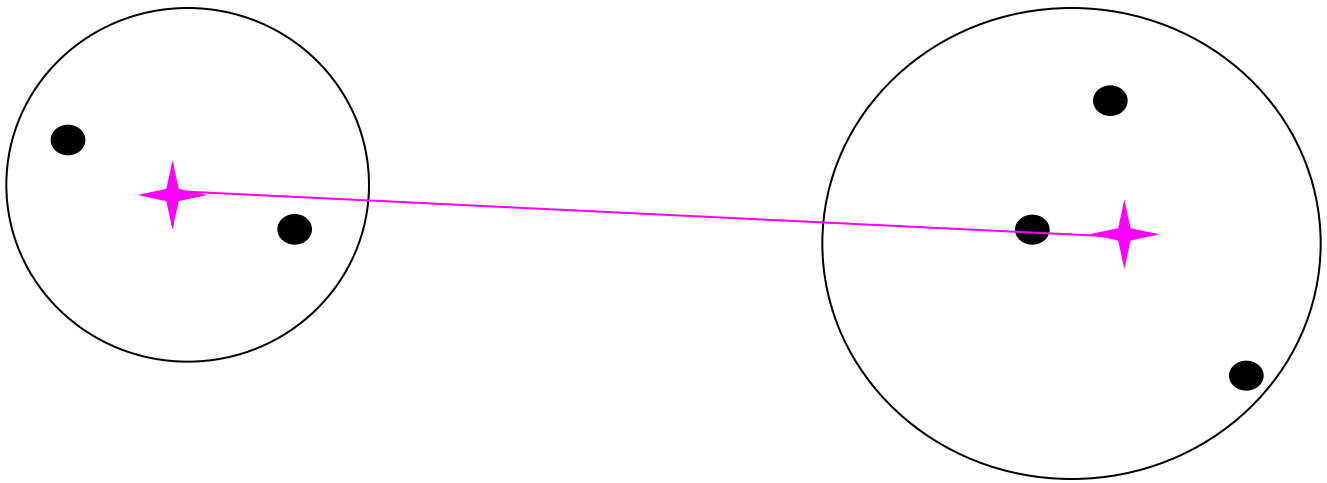


图 最短距离聚类法的过程

重心法

- 重心法（centroid hierarchical method）
  - 类的重心之间的距离
  - 对异常值不敏感，结果更稳定



重心法

4. 重心法

重心法定义类间距离为两类重心（各类样品的均值）的距离。重心指标对类有很好的代表性，但利用各样本的信息不充分。

设G_p与G_q分别有样品n_p，n_q个，其重心分别为$\bar{X}_p$和$\bar{X}_q$，则G_p与G_q之间的距离定义为$\bar{X}_p$和$\bar{X}_q$之间的距离，这里我们用欧氏距离来表示，即

$$D_{pq}^2 = (\bar{X}_p - \bar{X}_q)'(\bar{X}_p - \bar{X}_q)$$

重心法

- 设将G_p和G_q合并为G_r，则G_r内样品个数为$n_r = n_p + n_q$，它的重心是$\overline{X}_r = \frac{1}{n_r}(n_p\overline{X}_p + n_q\overline{X}_q)$，类$G_k$的重心是$\overline{X}_k$，那么依据（5.17）式它与新类$G_r$的距离为

$$D_{kr}^2 = \frac{n_p}{n_r}D_{kp}^2 + \frac{n_q}{n_r}D_{kq}^2 - \frac{n_p n_q}{n_r^2}D_{pq}^2$$

这里我们应该注意，实际上（**5.18**）式表示的类G_k与新类G_r的距离为：

-

重心法

$$D_{kr}^2 = (\bar{X}_k - \bar{X}_r)'(\bar{X}_k - \bar{X}_r)$$

$$= [\bar{X}_k - \frac{1}{n_r}(n_p\bar{X}_p + n_q\bar{X}_q)]'[\bar{X}_k - \frac{1}{n_r}(n_p\bar{X}_p + n_q\bar{X}_q)]$$

$$= \bar{X}_k'\bar{X}_k - 2\frac{n_p}{n_r}\bar{X}_k'\bar{X}_p - 2\frac{n_q}{n_r}\bar{X}_k'\bar{X}_q$$

$$+ \frac{1}{n_r^2}(n_p^2\bar{X}_p'\bar{X}_p + 2n_pn_q\bar{X}_p\bar{X}_q + n_q^2\bar{X}_q'\bar{X}_q)$$

重心法

- 利用 $\bar{X}_k'\bar{X}_k = \frac{1}{n_r}(n_p\bar{X}_k'\bar{X}_k + n_q\bar{X}_k'\bar{X}_k)$ 代入上式，有

$$D_{kr}^2 = \frac{n_p}{n_r}(\bar{X}_k'\bar{X}_k - 2\bar{X}_k'\bar{X}_p + \bar{X}_p'\bar{X}_p)$$

$$+\frac{n_q}{n_r}(\bar{X}_k'\bar{X}_k - 2\bar{X}_k'\bar{X}_q + \bar{X}_q'\bar{X}_q)$$

$$-\frac{n_p n_q}{n_r}(\bar{X}_p'\bar{X}_p - 2\bar{X}_p'\bar{X}_q + \bar{X}_q'\bar{X}_q)$$

$$= \frac{n_p}{n_r}D_{kp}^2 + \frac{n_q}{n_r}D_{kq}^2 - \frac{n_p n_q}{n_r^2}D_{pq}^2$$

重心法

- 【例】针对例5.1的数据，试用重心法将它们聚类。

（1）样品采用欧氏距离，计算样品间的平方距离阵$D^2_{(0)}$，见下表所示。

	G_1	G_2	G_3	G_4	G_5	G_6
G_1	0					
G_2	1	0				
G_3	16	9	0			
G_4	36	25	4	0		
G_5	64	49	16	4	0	
G_6	81	64	25	9	1	0

表

重心法

（2） $D^2_{(0)}$ 中最小的元素是$D^2_{12} = D^2_{56} = 1$，于是将G_1和G_2合并成G_7，G_5和G_6合并成G_8，并计算新类与其它类的距离得到距离阵$D^2_{(1)}$，见表

	G_1	G_2	G_3	G_4
G_1	0			
G_2	12.25	0		
G_3	30.25	4	0	
G_4	64	20.25	6.25	0

$$D^2_{37} = \frac{1}{2}D^2_{31} + \frac{1}{2}D^2_{32} - \frac{1}{2}\cdot\frac{1}{2}D^2_{12}$$

其中，

$$\frac{1}{2}\times 16 + \frac{1}{2}\times 9 - \frac{1}{2}\cdot\frac{1}{2}\times 1 = 12.25$$

其它结果类似可以求得

重心法

（3）在$D^2_{(1)}$中最小值是$D^2_{34}=4$，那么G_3与G_4合并一个新类G_9，其与与其它类的距离$D^2_{(2)}$，见表：

	G_7	G_9	G_8
G_7	0		
G_9	20.25	0	
G_8	64	12.5	0

表

重心法

（4）在中最小值是 = 12.5，那么与合并一个新类，其与与其它类的距离，见表：

	G_7	G_{10}
G_7	0	
G_{10}	39.0625	0

重心法

(5) 最后将G_7和G_{10}合并成G_{11}，这时所有的六个样品聚为一类，其过程终止。

上述重心法聚类的可视化过程见下图所示，横坐标的刻度表示并类的距离。

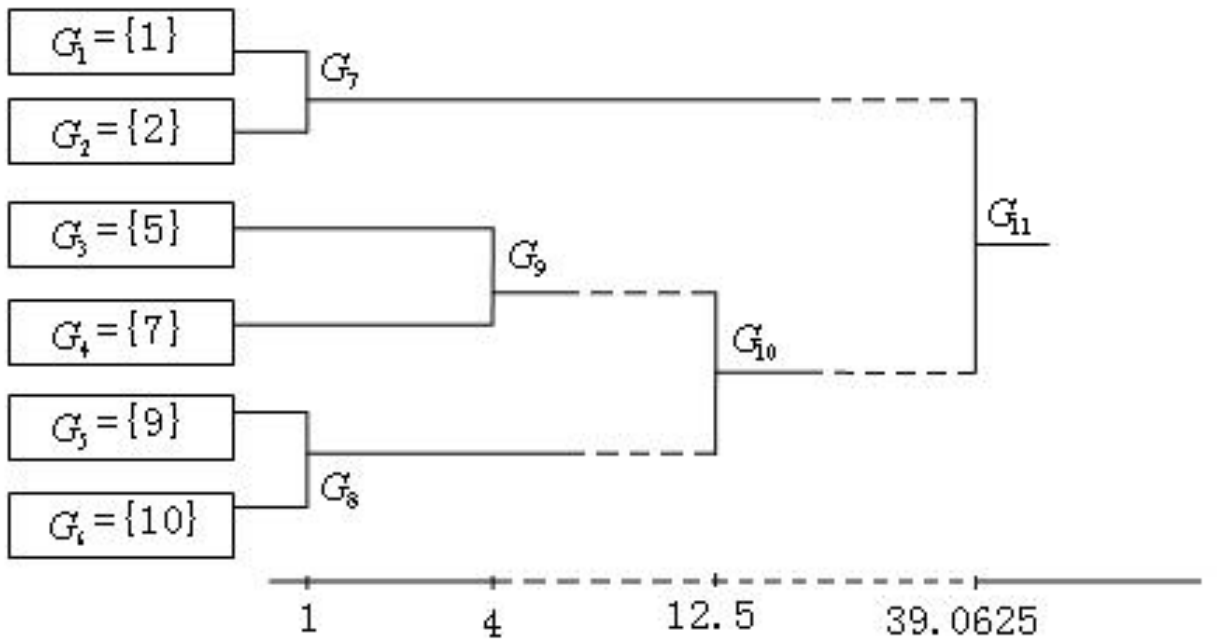


图 重心聚类法的过程

系统聚类法参数表

方　法	α_p	α_q	β	γ
最短距离法	1/2	1/2	0	-1/2
最长距离法	1/2	1/2	0	1/2
中间距离法	1/2	1/2	-1/4	0
重心法	n_p/n_r	n_q/n_r	$-\alpha_p\alpha_q$	0
类平均法	n_p/n_r	n_q/n_r	0	0
可变类平均	$(1-\beta)n_p/n_r$	$(1-\beta)n_q/n_r$	$\beta(<1)$	0
可变法	$(1-\beta)/2$	$(1-\beta)/2$	$\beta(<1)$	0
离差平方和	$(n_p+n_k)/(n_r+n_k)$	$(n_q+n_k)/(n_r+n_k)$	$-n_k/(n_k+n_r)$	0

层次聚类方法

层次的方法缺陷一旦一个步骤（合并或分裂）完成，就不能被撤销或修正，因此产生了改进的层次聚类方法，如

- BIRCH(balanced iterative reducing and clustering using hierarchies)算法
- CURE(clustering using representatives)算法
- ROCK(robua clustering using links)算法等

基于密度的方法
(density-based method)

- 主要有DBSCAN，OPTICS法
- **思想：**
  - 只要临近区域的密度超过一定的阈值，就继续聚类
- **特点：**
  - 可以过滤噪声和孤立点outlier，发现任意形状的类

基于密度的方法
(density-based method)

- 以空间中的一点为中心，单位体积内点的个数称为该点的密度。基于密度的聚类（density-basedclustering）根据空间密度的差别，把具有相似密度的相邻的点作为一个聚类。密度聚类只要邻近区域的密度(对象或数据点的数目)超过某个阈值，就能够继续聚类。
- 也就是说，对给定类中的每个数据点，在一个给定的区域内必须至少包含某个数目的点。这样，密度聚类方法就可以用来过滤“噪声”异常点数据，发现任意形状的簇。
- 在密度聚类算法中，有基于高密度连接区域的DBSCAN(Density-based Spatial Clustedng ofApplication with Noise)算法、通过对象排序识别聚类结构的OPTICS(Ordering Points To Identify the Clustering Structure)算法和基于密度分布函数聚类的DENCLUE(DENsity ．based CLUstEring)算法。

七、粒度计算

- 粒度计算从广义上来说是一种看待客观世界的世界观和方法论。
- 粒度计算的基本思想就是使用粒而不是对象为计算单元，使用粒、粒集以及粒间关系进行计算或问题求解。

粒度计算

- 1997年Lotfi A. Zadeh 提出了粒度的概念，他认为在人类认知中存在三种概念：粒度，组织与因果关系。从直观的来讲，粒化涉及到从整体到部分的分解，而组织却是从部分到整体的集成，而因果关系涉及原因与结果之间的联系。对一个事物的粒化就是以可分辨性、相似性、邻近性与功能性集聚有关的事物。
- 粒度计算是信息处理的一种新的概念和计算范式，覆盖了所有有关粒度的理论、方法、技术和工具的研究，主要用于处理不确定的、模糊的、不完整的和海量的信息。粗略地讲，一方面它是模糊信息粒度理论、粗糙集理论、商空间理论、区间计算等的超集，另一方面是粒度数学的子集。具体地讲，凡是在分析问题和求解问题中，应用了分组、分类、聚类以及层次化手段的一切理论与方法均属于粒度计算的范畴。信息粒度在粒度计算，词计算，感知计算理论和精化自然语言中都有反映

粒度计算的必要性

- **从问题求解的角度看**

  用粒度计算的观点来分析解决问题显得尤为重要，这样就不用局限于具体对象的细节。除此之外，将复杂问题划分为一系列更容易管理和更小的子任务，可以降低全局计算代价。

- **从应用技术的角度看**

  图像处理、语音与字符识别等，是计算机多媒体的核心技术。这些信息处理质量的好坏直接依赖于分割的方法和技术，而粒度计算的研究或许能够解决这一问题。

粒度计算的基本问题

- 两大问题
  - 粒的构造 ：处理粒的形成、表示和解释
  - 使用粒的计算：处理在问题求解中粒的运用
- 两个方面
  - 从语义 上：侧重于对粒的解释， 如为什么两个对象会在同一个粒之中， 为什么不同的粒会相关。
  - 从算法上：如何进行粒化和如何进行基于粒的计算。对粒的分解与合并方法的研究，是构建任何粒度体系结构的本质要求。

粒度计算的国内外研究现状

- 粗糙集理论
  - 粒：等价类，子集
  - 粒的计算：粒之间的近似
- 商空间理论
  - 粒：等价类，子集，粒之间具有拓扑关系
  - 粒的计算：合成、分解
- 词计算理论
  - 粒：词
  - 粒的计算：模糊数学

文献及其他资源

- F. Sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys, 34(1): pp. 1-47, 2002.
- Li J Y, Sun MS, Zhang X. A comparison and semi-quantitative analysis of words and character-bigrams as features in Chinese text categorization. COLING-ACL’ 06
- Pu Wang, Carlotta Domeniconi. Building Semantic Kernels for Text Classification using Wikipedia. KDD 08’
- Xuan-Hieu Phan,Le-Minh Nguyen, Susumu Horiguchi. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. WWW’ 08
- W.Y. Dai, G.R. Xue, Q. Yang and Y. Yu, Transferring Naive Bayes Classifiers for Text Classification, AAAI 07’
- C.Do, A. Ng, Transfer Learning for text classification. NIPS’ 05
- F. Mourão, L. Rocha, et al., Understanding Temporal Aspects in Document Classification, WSDM 07’

NEXT：对话生成与问答系统

- 9月23日：对话生成与问答系统
- 10月12日补课(双周周一课程)：topic models ;)
  - Dr. Jianhua Yin