



第六章：

话题检测跟踪技术

杨建武

北京大学计算机科学技术研究所

Email: yangjianwu@icst.pku.edu.cn

话题检测跟踪技术



- Topic Detection and Tracking (TDT)
- 话题检测跟踪技术
- 主题检测与追踪技术
- TDT technique explore for detecting the appearance of **new topics** and for **tracking** the reappearance and evolution of them.



相关概念

术语



➤ Event (事件)

- ❖ A specific thing that happens at a specific **time** and **place** along with all **necessary preconditions** and **unavoidable consequences**.
- ❖ Specific elections, accidents, crimes and natural disasters are examples of events. Eg. 911事件

➤ Activity (活动)

- ❖ An activity is a **connected set of actions** that have a common focus or purpose.
- ❖ Specific campaigns, investigations, and disaster relief efforts are examples of activities.
- ❖ Eg. 党的十七大、北京奥运会

术语



➤ **Topic**（话题、主题）

- ❖ an event or activity, **along with** all directly related events and activities
- ❖ Topic \rightarrow Event, $1 \rightarrow n, n \geq 1$
- ❖ Sometimes, Topic \approx Event (99年前)

➤ **Topic Example**

- ❖ **WHAT**: 35 or 40 young Mountain Hikers were lost in an avalanche in France around the 20th of January.
- ❖ **WHERE**: Orres, France
- ❖ **WHEN**: January 4, 1998



➤ Story (报道)

- ❖ A story is a newswire **article** or a segment of a news broadcast with a coherent news focus.
- ❖ They must contain at least two independent, declarative clauses (子句).



➤ Story Segmentation

- ❖ dividing the transcript of a news show into **individual stories**

➤ First Story Detection

- ❖ recognizing the **onset (开始)** of a **new topic** in the stream of news stories

➤ Story Link Detection

- ❖ deciding whether two randomly selected stories **discuss the same news topic**



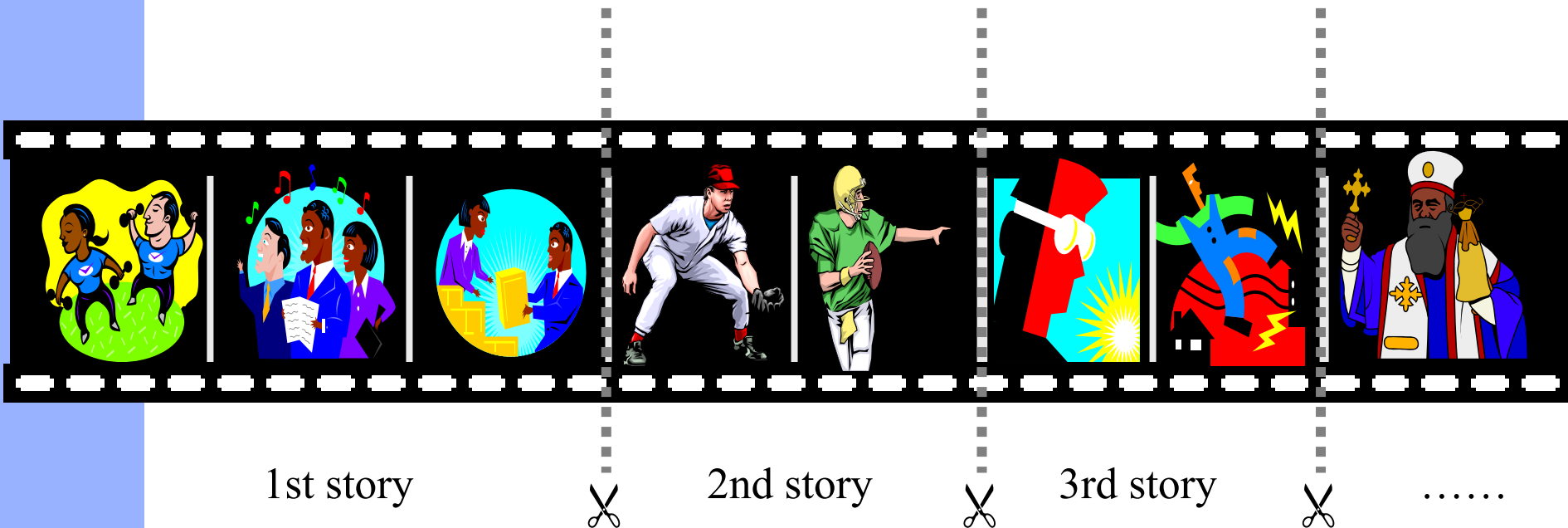
➤ Topic Detection

❖ **grouping** all stories as they arrive, based on the **topics** they discuss

➤ Topic Tracking

❖ monitoring the stream of news stories to **find additional stories** on a **topic** that was identified using several sample stories

Story Segmentation



Story Segmentation



➤ goal

- ❖ take a show of news and to detect the **boundaries** between stories automatically

➤ types

- ❖ done on the **audio** source directly
- ❖ using a text **transcript** of the show—
either closed captions or speech
recognizer output

Story Segmentation



➤ approaches

- ❖ look for **changes** in the vocabulary that is used
- ❖ look for words, phrases, pauses, or other features that occur near story boundaries, to see if they can find sets of features that reliably **distinguish** the middle of a story from its beginning or end, and clustering those segments to find larger story-like units

First Story Detection



➤ goal

- ❖ recognize when a news topic appears that had **not been discussed earlier**
- ❖ Detect that **first** news story that reports a bomb's explosion, a volcano's eruption, or a brewing political scandal

➤ applications

- ❖ interest to information, security, or stock analysts whose job is look for **new events** that are of significance in their area

First Story Detection



➤ approach

- ❖ (1) Reduce stories to a set of **features**, either as a vector or a probability distribution.
- ❖ (2) When a new story arrives, its feature set is **compared** to those of *all* past stories.
- ❖ (3) If there is sufficient **difference** the story is marked as a first story; otherwise, not.

Story Link Detection



➤ goal

❖ handed two news stories, determine whether or not they discuss the same topic

今天上午近百位達悟人又回到蘭嶼核廢料貯存場，聚集在核廢桶儲存場的草原上，等待高金素梅等..

林義夫今天上午前往蘭嶼與當地居民面對面溝通，並達成核廢場址遷移的共識，傍晚返回台北，在行政..

呂秀蓮針對台灣核廢料送往大陸處理的問題指出，如果大陸當局確實願意伸出援手，協助台灣解決..

The ruling party's committee on reform of the legislature supports a reduction in the number

Taiwan should not to "push too hard" in capitalizing on the current good relations with the US government, a US scholar.

世界盃足球賽即將在五月底在日本和韓國揭幕，C組的中國大陸代表隊將在六月四日迎戰哥斯大黎加，哥

Officials say that the nation's water resources should keep until the end of June; and if it rains a little this month, the

"The dam's dead storage amounts to about 15,000 tonnes of water, which is enough to support us until the end of June," said Kuo Yao-chi (郭瑤琪), executive-general of the center.

Yes / No

Yes / No

Yes / No

Yes / No

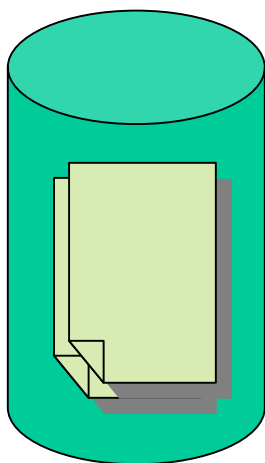
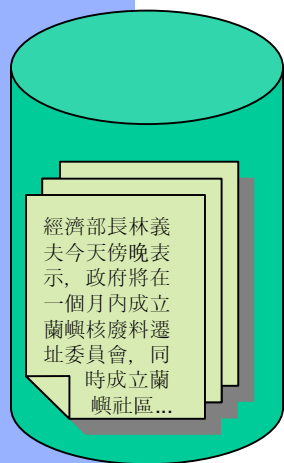
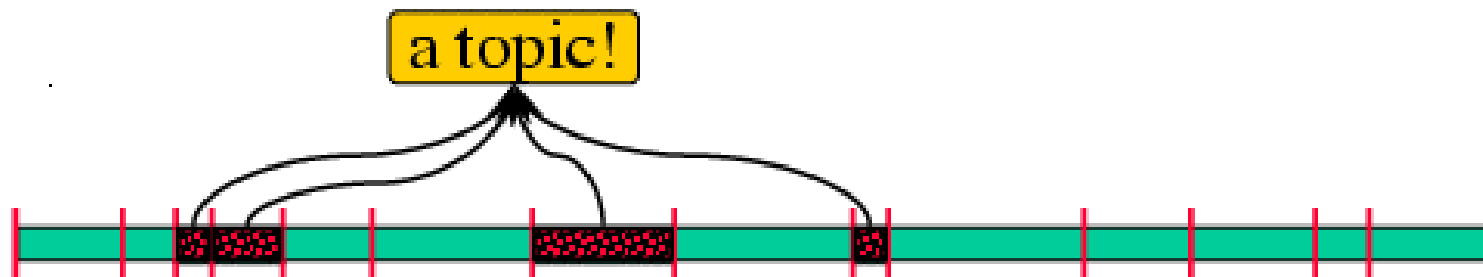
Topic Detection



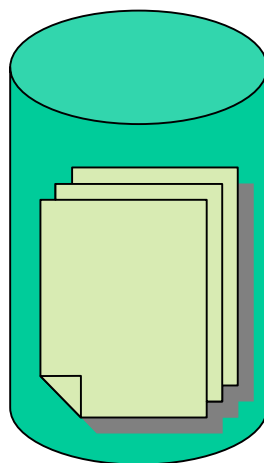
➤ goal

- ❖ to cluster stories on the **same** topic into bins
- ❖ the creation of bins is an **unsupervised** task

Topic Detection



.....



Taiwan should not to "push too hard" in capitalizing on the current good relations with the US government, a US scholar.

世界盃足球賽即將在五月底在日本和韓國揭幕，C組的中國大陸代表隊將在六月四日迎戰哥斯大黎加，哥

The ruling party's committee on reform of the legislature supports a reduction in the number

今天上午近百位達悟人又回到蘭嶼核廢料貯存場，聚集在核廢桶儲存溝的草原上，等待高金素梅等..

Officials say that the nation's water resources should keep until the end of June; and if it rains a little this month, the

陳水扁總統今天上午在總統府接見第九屆十大傑出愛心媽媽，對她們無私的奉獻，表達感佩之意...

.....

Topic Detection



➤ approach

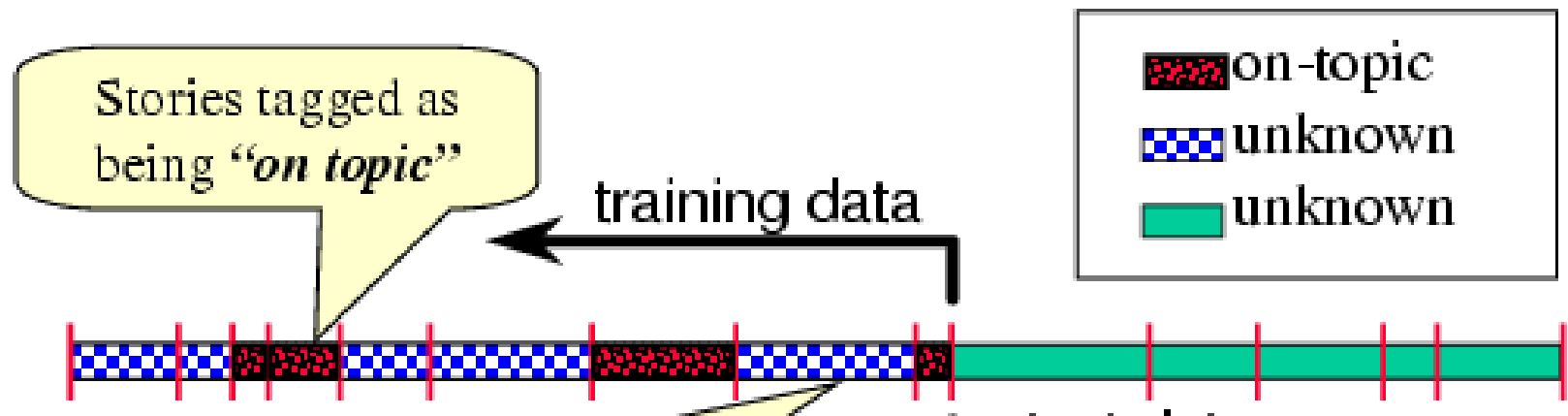
- ❖ (1) Stories are represented by a set of **features**.
- ❖ (2) When a new story arrives it is compared to all past stories and assigned to the cluster of the **most similar story** from the past (i.e., one nearest neighbor).



Topic Tracking

➤ goal

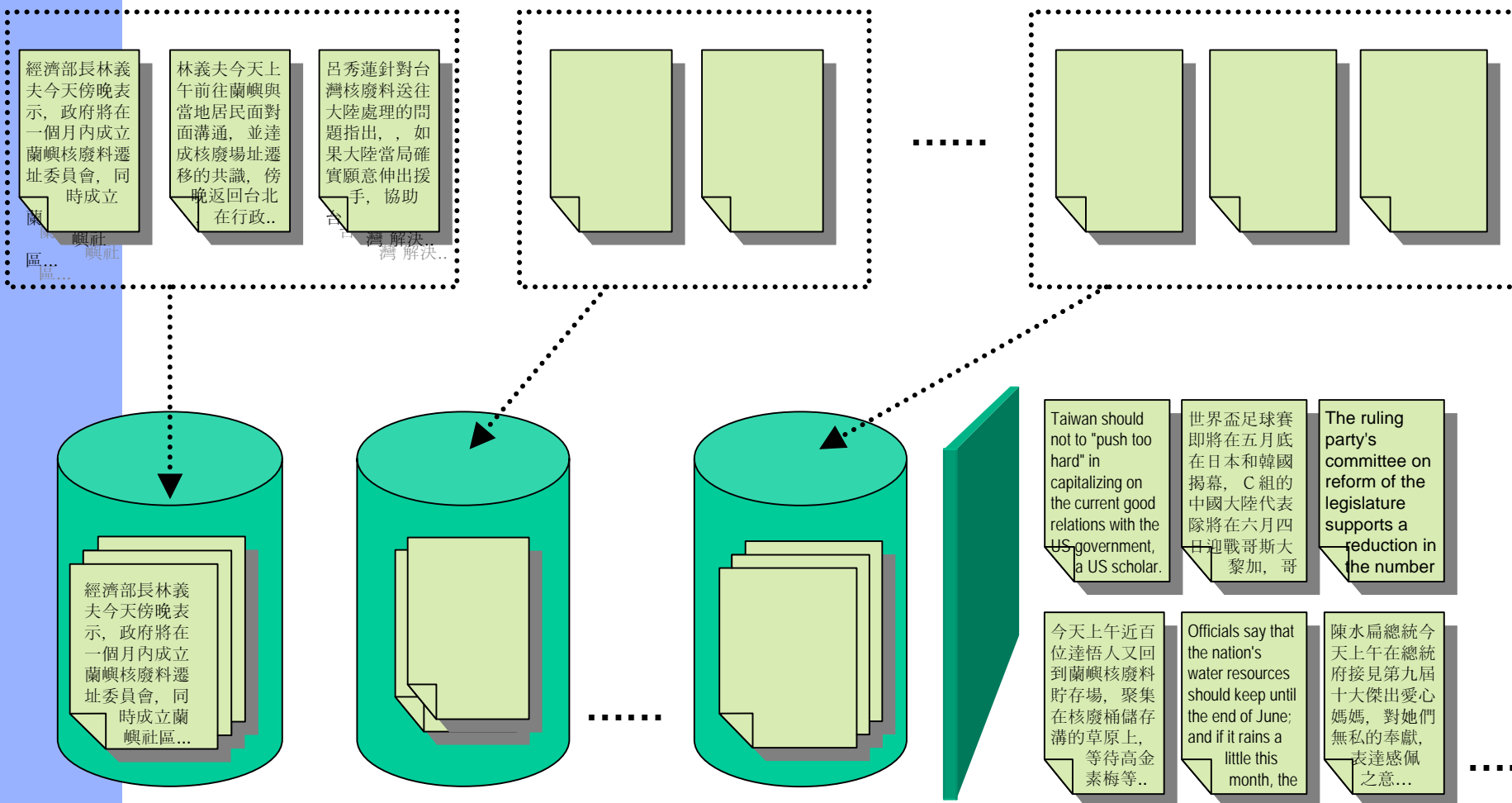
- ❖ similar information retrieval's filtering task
- ❖ provided with a small number of stories that **are known** to be on the same topic, find all other stories on that topic in the stream of **arriving** news





Topic Tracking

documents of
the same topic



Topic Tracking



➤ approach

- ❖ extract a set of features from the **training stories** that differentiate it from the much larger set of stories in the past
- ❖ When a new story arrives, it is compared to the topic features and if it **matches** sufficiently, declared to be on topic.



应用

TDTLightHouse



- James Allan, University of Massachusetts
- 马萨诸塞大学Amherst分校智能信息检索中心
- TDTLightHouse系统建立在交互式信息检索系统LightHouse系统之上
- TDT核心系统则运行在后台
- 对模拟的新闻信息流进行话题检测，新闻信息流按照话题形成一个个话题类簇
- TDTLightHouse的客户端实时查询检测结果，并以图示化的界面展现给用户



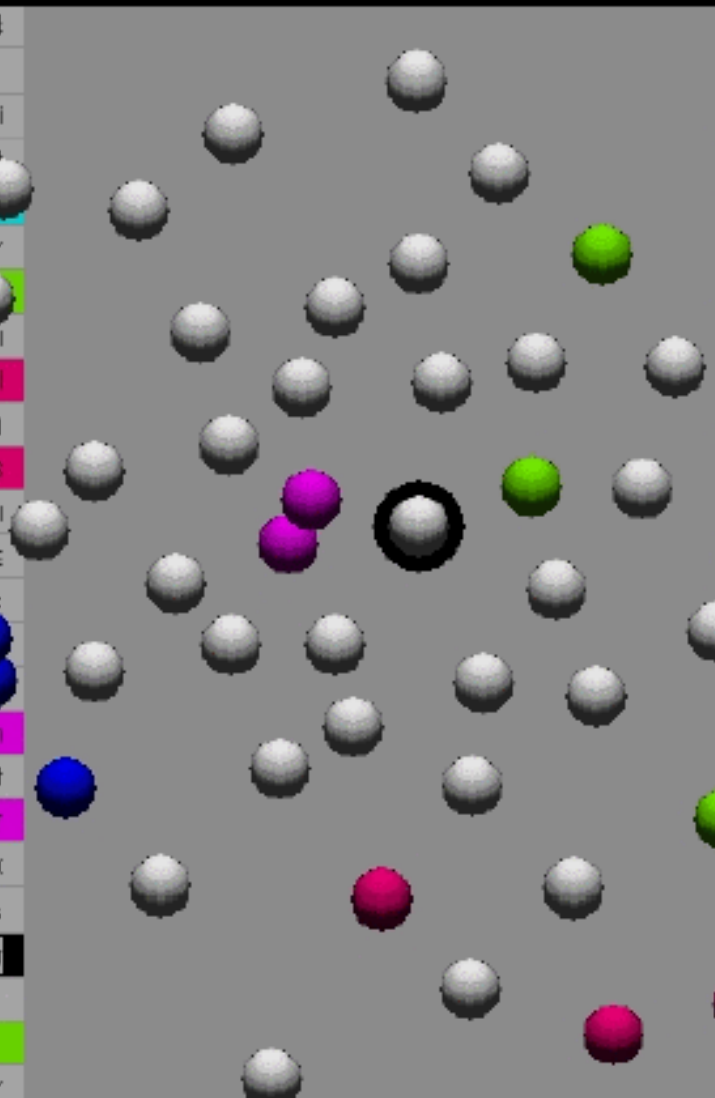
Query:

1

TDT Engine:

UMass

- ☐ 1. tile tiles party glass Party Republ
- ☐ 2. Turkey Syria Turkish Kurdish m
- ☐ 3. Taiwan China Koo Chinese Bei
- ☐ 4. Pinochet Chile Chilean Spanist
- ☐ 5. House president Clinton imp
- ☐ 6. gay Shepard people Wyoming I
- ☐ 7. weather -- people forecast
- ☐ 8. site Web sites news information
- ☐ 9. Yankees series game on I
- ☐ 10. Il boundaries separate -- new g
- ☐ 11. match set 6-3 Sampras Open s
- ☐ 12. witches books witchcraft Wicca
- ☐ 13. Goldman Russia Russian gove
- ☐ 14. Iraq weapons United Iraqi sanc
- ☐ 15. VOA News Tom Susan Cro
- ☐ 16. VOA News re hours news II
- ☐ 17. Brazil Cardoso percent govern
- ☐ 18. people News Tonight good nigt
- ☐ 19. world financial IMF Clinton ecor
- ☐ 20. care Medicare health HMOs per
- ☐ 21. Jones Clinton president case s
- ☐ 22. Kosovo NATO Milosevic ethnic i
- ☐ 23. war Cohen family Bosnian Bell
- ☐ 24. rain weather northern showers
- ☐ 25. Taliban Iran Afghanistan Irania



- ☐ 26. Microsoft Netscape governmen
- ☐ 27. Japan yen Japanese dollar per
- ☐ 28. plan problems officials nations
- ☐ 29. president House Clinton White
- ☐ 30. Wall Street points Dow today m
- ☐ 31. card credit Internet Mastercard c
- ☐ 32. Arab million grain tons team Gr
- ☐ 33. Strawberry cancer colon Yanker
- ☐ 34. World 91 PRI re Radio Public W
- ☐ 35. hurricane Mitch miles people M
- ☐ 36. Ha ... News -- two minutes i
- ☐ 37. Amato Schumer New York cam
- ☐ 38. stock Internet companies Singa
- ☐ 39. Kong Hong China government i
- ☐ 40. Canada seats council Greece b
- ☐ 41. Israeli Palestinian Netanyahu Is
- ☐ 42. Korea North South Korean Japa
- ☐ 43. pl ... owners NBA union sea:
- ☐ 44. Germany Schroeder governmer
- ☐ 45. Glenn space John shuttle flight
- ☐ 46. percent points market stocks D
- ☐ 47. rebels Congo government troop
- ☐ 48. Russia Yeltsin Russian govern
- ☐ 49. prize peace Nobel work Ireland
- ☐ 50. voice actors New energy years 1



热点信息

综合 新闻网站 新闻

当前时段：最新24小时 选择时段：最新24小时

查看热

温家宝抵达日本“融冰之旅”起航(图) 国际

详细>

1

千龙新闻网 (2007-04-12 08:10:00)

内容提要：这是中国总理7年来首次访日。温家宝还将在日本国会发表演讲。他说。为在亚洲和世界上具有重要影响的国家，中日两国之间的关系是最重要的双边关系之一。

21CN- 白鸽网- 北方网- 所有151条相关 - 17条评论

美专家：伊朗是在“吹牛” 国际

详细>

2

东方圣域网 (2007-04-12 08:13:00)

内容提要：这是4月9日拍摄的伊朗纳坦兹核设施一角。在谴责伊朗的同时，西方国家也呼吁伊朗进行谈判。迈克尔说：“从政治上来讲，拥有3000台离心机比正确运行这些离心机更重要...”

TOM- 百灵网- 半岛都市报- 所有118条相关 - 2条评论

中国成功发射第二颗海洋卫星海洋一号(图) 军事

详细>

3

千龙新闻网 (2007-04-12 07:15:00)

内容提要：4月11日11时27分，我国自行研制的“海洋一号B”卫星在太原卫星发射中心发射升空。作为“海洋一号A”卫星的后续星，“海洋一号B”卫星是中国海洋立体监测系统的重要组成部分...

北方网- 长城在线- 东北新闻网- 所有73条相关 - 87条评论

阿尔及利亚首都阿尔及尔发生数起爆炸 国际

详细>

4

中国网 (2007-04-12 08:07:00)

<<

日 一

1 2

8 9

15 16

22 23

29 30

前五热



日本 伊朗



结果评价



Evaluation Metrics

- Miss: $\text{miss} = c/(a+c)$
- False alarm (f): $\text{fa} = b/(b+d)$
- Recall (r) : $r = a/(a+c)$
- Precision (p) : $p = a/(a+b)$
- F1 measure(F1), micro、 macro
- $F1 = 2rp/(r+p)$

	In topic	Not in topic
In topic (system)	a	b
Not in topic (system)	c	d



Cost Functions

$$C_{Det} = C_{Miss} \cdot P_{Miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{non-target}$$

$$(C_{Det})_{norm} = C_{Det} / \text{MIN}(C_{Miss} \cdot P_{target}, C_{FA} \cdot P_{non-target})$$

C_{Miss} (e.g., 10) and C_{FA} (e.g., 1) are the costs of a missed detection and a false alarm respectively, and are **pre-specified** for the application.

P_{Miss} and P_{FA} are the probabilities of a **missed detection** and a **false alarm** respectively and are determined by the evaluation results.

P_{Target} is the **a priori probability** (先验概率) of finding a target as specified by the application.

$$P_{non-Target} = 1 - P_{Target}$$



技术方法

发展历程



➤ Begin at 1998

❖ J. Allan, R. Papka, and V. Lavrenko.

On-line new event detection and tracking.

In Proc. of SIGIR Conference on Research and Development in Information Retrieval, 1998

❖ Y. Yang, T. Pierce, and J. G. Carbonell.

A study on retrospective and on-line event detection.

In Proc. of SIGIR Conference on Research and Development in Information Retrieval, 1998

发展历程



❖ 提出问题，并给出解决方案

- the similarities between the **incoming documents** and the **known events** (sometime represented by a **centroid**) are computed, and then a threshold is applied to make decision whether the incoming document is the first story of a **new** event or a story of some **known event**.

发展历程



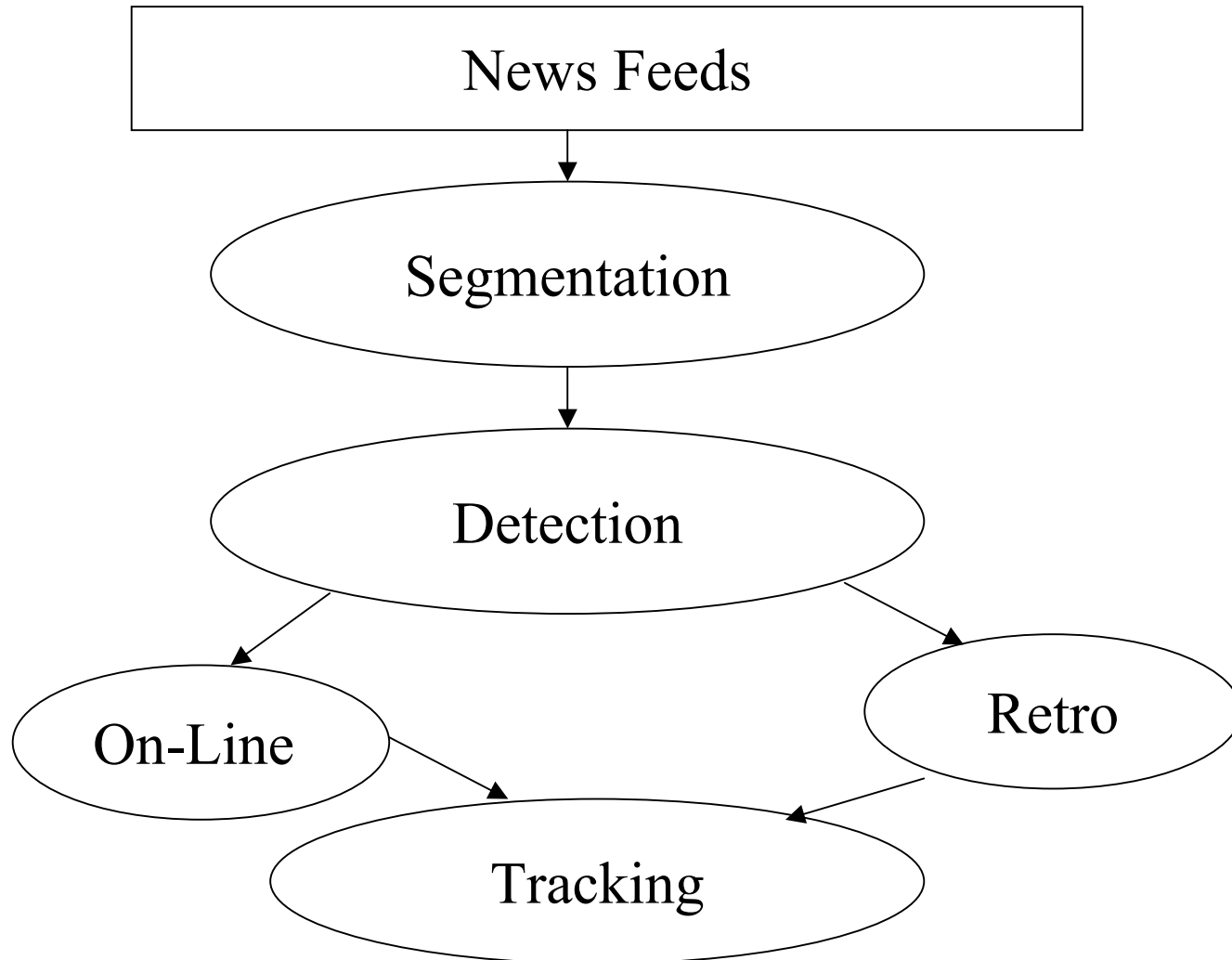
➤ Milestone papers

- ❖ J. Allan, H. Jin, M. Rajman, C. Wayne, G. D., L. V., R. Hoberman, and D. Caputo. Summer workshop final report. In Center for Language and Speech Processing, 1999.
- ❖ W. Lam, H. Meng, K. Wong, and J. Yen.
Using contextual analysis for news event detection. International Journal on Intelligent Systems, 2001.
- ❖ Y. Yang and J. Z. et al. Topic-conditioned novelty detection. In Proc. of the **SIGKDD** international conference on Knowledge discovery and data mining, 2002.
- ❖ T. Brants, F. Chen, and A. Farahat. A system for new event detection. In Proc. of the **SIGIR** conference on Research and development in information retrieval, 2003.
- ❖ Zhiwei Li, Bin Wang, M-J. jing Li, W-Y Ma,
A probabilistic Model for retrospective news event detection. In Proc. of the **SIGIR** conference on Research and development in information retrieval, 2005.



话题检测

Tasks in News Detection





两种类型的话题检测

➤ NED&RED

❖ NED

- New Event Detection
- labeling each document in an sequence stream with a New or Old Flag
- On-line

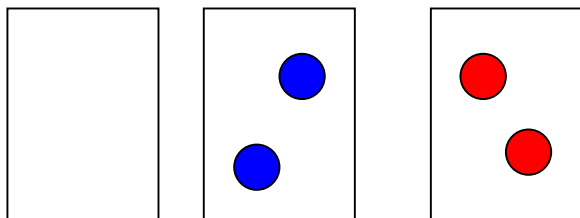
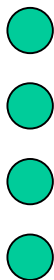
❖ RED

- Retrospective Event Detection (回溯)
- discovery of previously unidentified events in historical news corpus
- Off-line



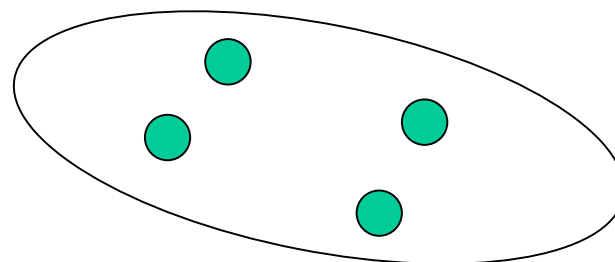
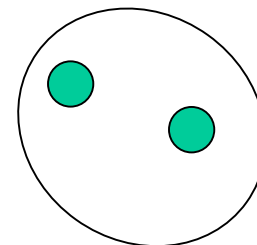
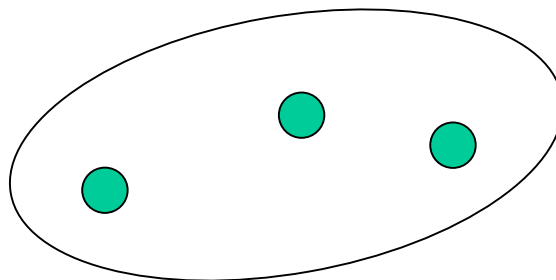
NED

document in
sequence stream



RED

historical news corpus



Deferral Period



- How long the system can **delay** when making a decision
- How many news articles the system can **look ahead**
- The “burst” nature of news articles
- The deferral period is defined in DEF
- **DEF = 10**

话题检测技术



- A study on retrospective and on-line event detection
- Two Methods
 - ❖ GAC-based hierarchical clustering,
GAC (Group Average Clustering)
基于平均分组的层次聚类法
 - ❖ Single-pass clustering

话题检测– GAC算法

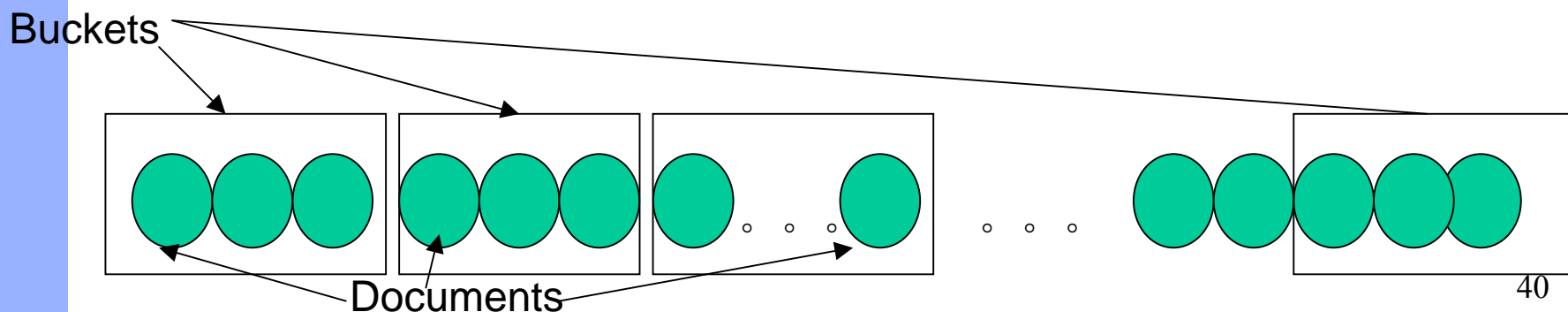


- 基于平均分组的层次聚类法（GAC）是针对回溯检测(RED)的一种较好算法
- 一种自底向上的贪心算法，采用了分而治之的策略
- 能够最大化话题类簇中各新闻报道之间的平均相似度
- 输入为按照时间排好序的新闻文档集合
- 输出为层次式的话题类簇结构。

话题检测– GAC算法



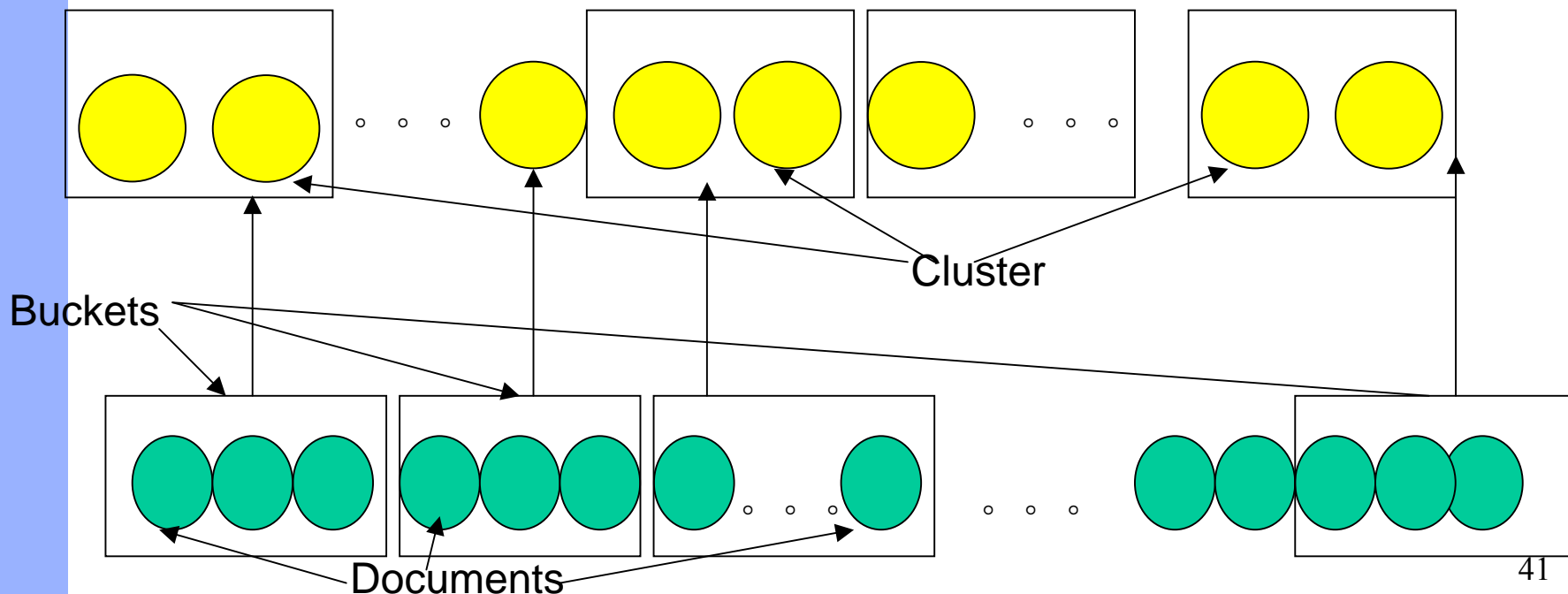
- 1) 初始将文档集合中**每篇**文档作为一个单独的话题类簇，
 - ❖ 初始划分为所有单篇文档组成的话题类簇；
- 2) 将当前话题类簇集合中的**话题类簇**按顺序**连续**并且不重叠地划分到**大小为m的桶**中；



话题检测– GAC算法



- 3) 对每一个桶分别进行聚类，
 - ❖ 重复地合并桶中两个最相似的低层次话题类簇，形成一个高层次的话题类簇，
 - ❖ 直到桶中类簇数量减少的比例达到预设的 p 为止，或者任何两个类簇之间的相似度值均低于一个预定义的阈值 s 为止。

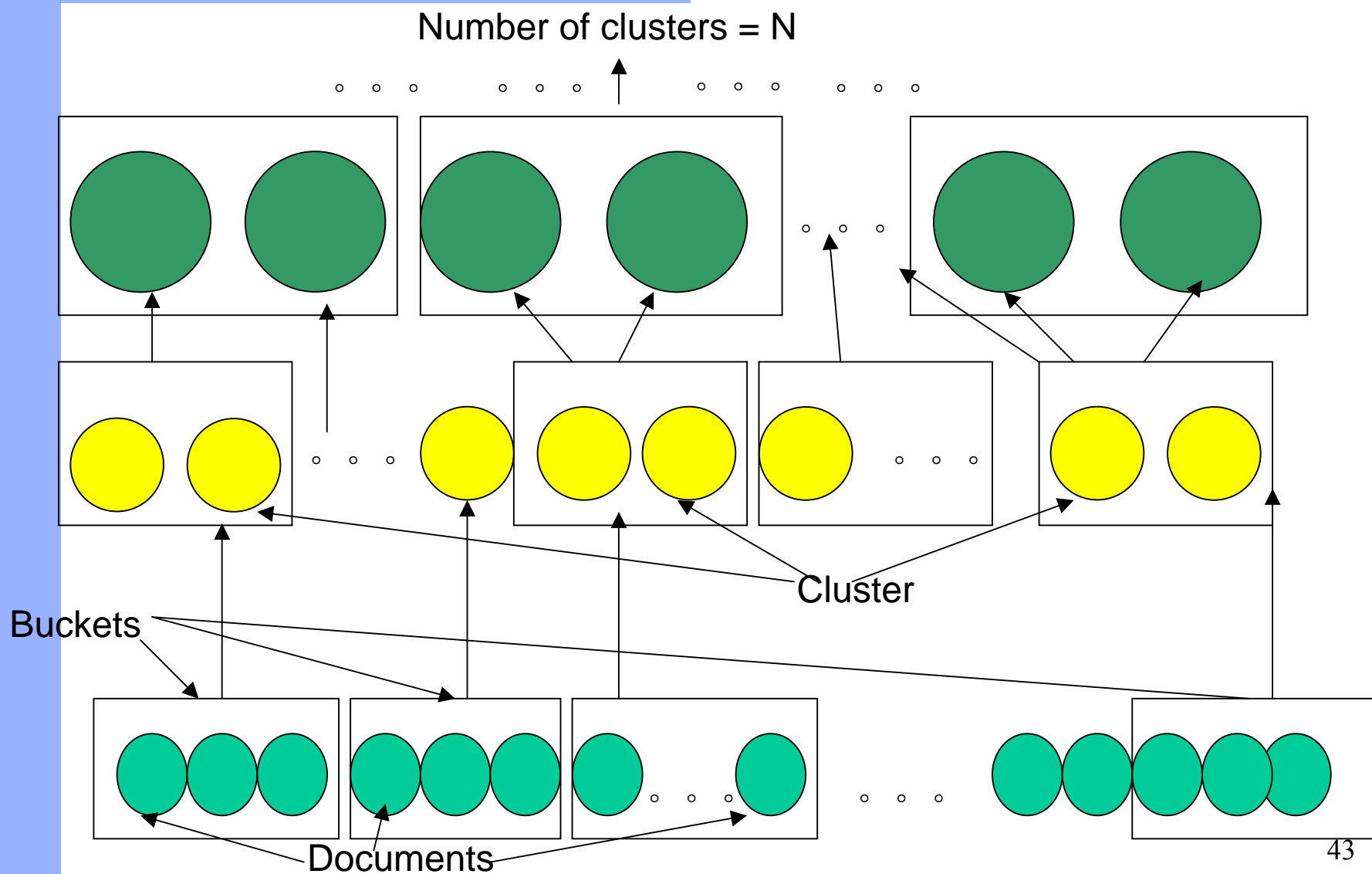


话题检测– GAC算法



- 4) 在保持各话题类簇时间顺序的前提下去除桶的边界，也即汇集所有桶中的话题类簇。此时对文档集合的划分即为当前**类簇集合**。

话题检测-GAC算法

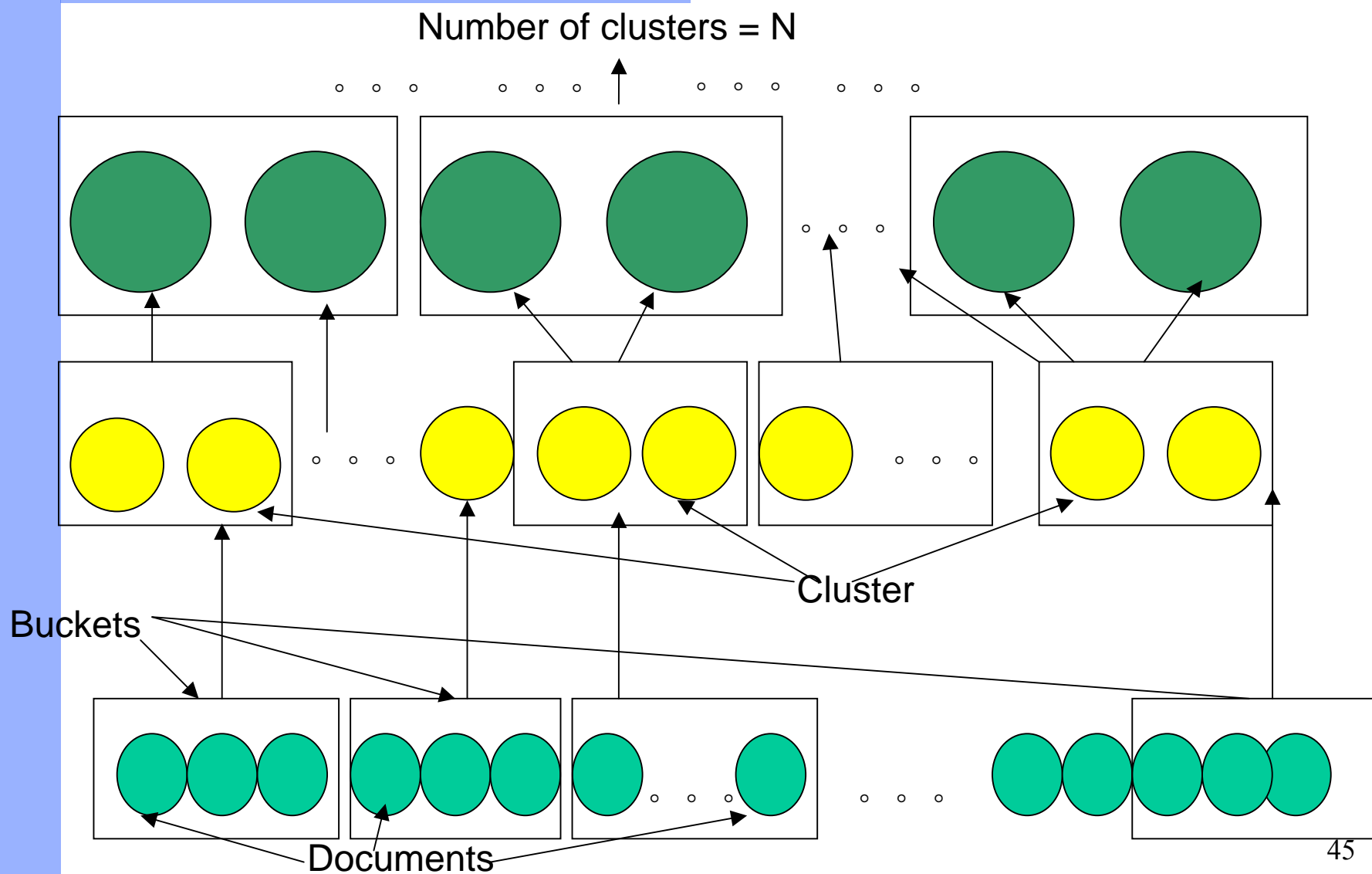


话题检测– GAC算法



- 5) 重复2)、3)、4)三步，直到最顶层的话题类簇数目达到了一个预定的数值为止；
- 6) 定期地将每个顶层类簇中的所有新闻文档按照前五步进行重新聚类。

话题检测-GAC算法





话题检测– GAC算法分析

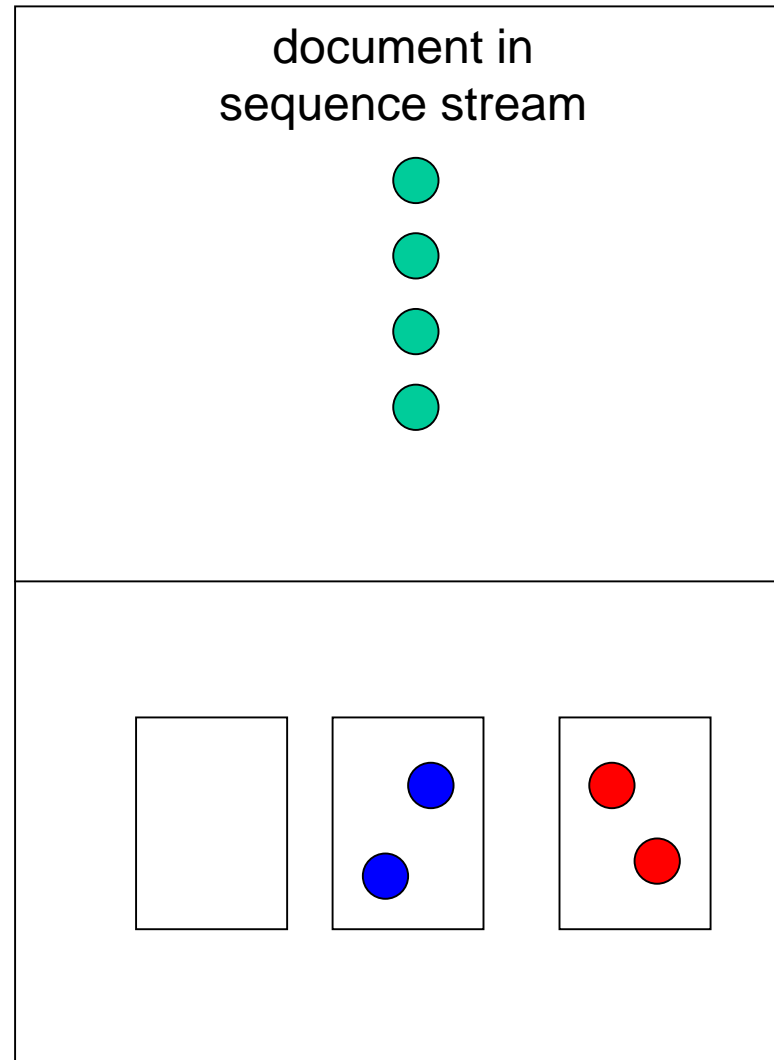
- GAC算法的**时间复杂度为 $O(mn)$**
其中 n 为新闻文档集合中的文档数量， m 为桶的大小， $m \leq n$ 。
- 该算法不仅效率高，而且通过考虑新闻文档的**时间顺序**提高了话题类簇的质量。
- 通过调整该算法中用到的参数可以改善检测结果。
- GAC算法**只适合回溯检测**，不适合话题的在线检测，因此应用范围受到了一定的限制。

话题检测–Single Pass Clustering



- Assign each **incoming** document to one of a set of topic clusters
- A topic cluster is represented by its **centroid** (vector average of members)
- For incoming story compute similarity with centroid

话题检测–Single Pass Clustering



话题检测–Single Pass Clustering



➤ 两个阈值:

❖ 聚类阈值 t_c 创新阈值 t_n , $t_c \geq t_n$ 。

➤ 算法描述: 当前新闻报道 x 与以前某个新闻话题 T 之间具有最大相似度值 $\text{sim}_{\max}(x)$ 。

1) if($\text{sim}_{\max}(x) > t_c$)

{ x 被标识为"OLD", 属于该话题 T ; }

2) if($t_n < \text{sim}_{\max}(x) \leq t_c$)

{ x 被标识为"OLD", 不处理; }

3) if($\text{sim}_{\max}(x) \leq t_n$)

{ x 被标识为"NEW", 创建一个新话题; }

话题检测–Single Pass Clustering



➤ 时间窗口

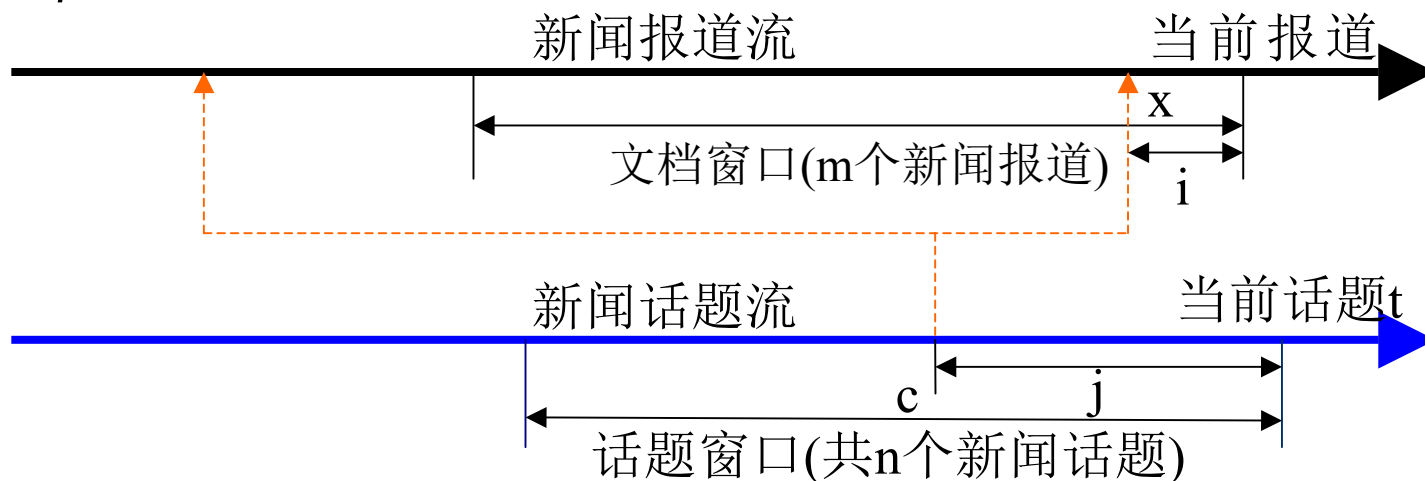
目的：利用属于同一话题的新闻报道之间的**时间相邻性**，影响当前文档 x 和话题 c 之间相似度值。



话题检测 – 时间窗口

$$sim(\vec{x}, \vec{c})'' = \begin{cases} (1 - \frac{i}{m} \cdot \alpha) \times (1 - \frac{j}{n} \cdot \beta) \times sim(\vec{x}, \vec{c}), & \text{如果 } c \text{ 中有文档在文档窗口中;} \\ 0, & \text{否则;} \end{cases}$$

- i 表示当前新闻文档 x 和话题类簇 c 中与 x 时间最相近的文档之间相隔的文档个数。当前新闻文档 x 之前的 m 个新闻文档
- j 表示话题 c 和话题窗口中最后一个话题之间相隔的话题个数
- $\alpha\beta$ 为调节因子。





IDF动态改变

- 新闻演化发展，使用的词汇在变化
- 自适应的IDF计算方法(CMU)

$$IDF(t, p) = \log_2 (N(p) / n(t, p))$$

❖ p表示当前时间点



IDF动态改变

- 话题类簇相关的IDF值(IBM)
- 初始IDF值为标准的与话题无关的 $idf_0(w)$
- 随着文档变化，IDF值不断调整

$$idf(w, cl) = idf_0(w) + \Delta idf(w, cl)$$

$$\Delta idf(w, cl) = \lambda \frac{2|D_w \cap cl|}{|D_w| + |cl|}$$

- ❖ λ 为调整因子;
- ❖ D_w 包含词 w 的文档集合, $|D_w|$ 文档数;
- ❖ 类簇 cl 中的文档集合, $|cl|$ 文档数;
- ❖ $|D_w \cap cl|$ 表示和 cl 交集集中的文档数目。

Named Entities



- Named entities, which denote people, places, time, events, and things, play an important role in a news story
- Solutions
 - ❖ Named Entities with Amplifying Weights **before** Selecting
 - ❖ Named Entities with Amplifying Weights **after** Selecting

Named Entities



Named Entities with Amplifying Weights **before Selecting**

amplification	P(miss)	P(FA)	Cdet (norm)
weight \times 1	0.4010	0.0060	0.4304
weight \times 2	0.4335	0.0038	0.4519
weight \times 3	0.4559	0.0032	0.4714

Named Entities with Amplifying Weights **after Selecting**

amplification	P(miss)	P(FA)	Cdet (norm)
weight \times 1	0.4010	0.0060	0.4304
weight \times 2	0.3630	0.0027	0.3763
weight \times 3	0.3552	0.0037	0.3740



阈值动态改变

- 基于时间的阈值模型（马萨诸塞州立大学）
- 时间上距离某个话题越远的新闻报道越难加入该话题

$$\begin{aligned} threshold(q_i, d_j) = & 0.4 + \theta * (sim(q_i, d_j) - 0.4) \\ & + \beta * (date_j - date_i) \end{aligned}$$

- ❖ $sim(q_i, d_j)$ 为当前新闻报道与话题类簇之间的相似度；
- ❖ $(date_j - date_i)$ 为新闻报道的到达时间与话题类簇创建时间之间间隔的天数；
- ❖ θ, β 为调整因子；
- ❖ 0.4 为 Inquiry[29] 系统中的经验值。



话题跟踪

话题跟踪



- 话题跟踪：监控新闻报道信息流以便发现与某一**已知话题**有关的新报道；
- 步骤：首先给出一组**样本**报道，训练得到一个话题模型，然后系统在后续报道中要能**识别**出其后关于此话题的所有报道。
- 实质：显然可以把话题跟踪看作是一种特殊的二元分类问题。
- **文本分类**技术是话题跟踪的基础

话题跟踪



➤ 特点:

- ❖ 面向动态的、随时间变化的新闻报道信息流，而不是静态的文本集合；
- ❖ 对新闻报道流进行实时跟踪，不能有延迟
- ❖ 相对于大量的未标注报道而言，可用于训练的正例数量非常有限。

➤ 常用算法:

- ❖ 基于查询的方法：构造查询向量
- ❖ 基于单元语言模型的方法
- ❖ 基于分类的方法：kNN、SVM



基于查询的话题跟踪

➤ 跟踪器构造：只利用相关报道。

❖ 两个阈值：

- 跟踪阈值 t_1 ；跟踪器调整阈值 t_2 , $t_1 < t_2$ 。

❖ 构造过程：

- 将训练集中出现的非停用词 w 按照其对应的 $r * idf(w)$ 值由高到低排序

其中 r 为包含词 w 的相关报道数量， $idf(w)$ 为词 w 的倒排文档频率；

- 取前 n 个词组成查询向量，
查询向量第 k 维(词 w)的取值 $q(k) = tf(w) * idf(w)$ ，
 $tf(w)$ 为所有相关报道中词 w 的平均 tf 值



基于查询的话题跟踪

- 新闻报道的特征向量为构建跟踪器时选择的 n 个词所组成
- 对应跟踪器中第 k 维(词 w)，新闻报道向量的第 k 维(词 w) 的取值：

$$dk=0.4+0.6*tf(w)*idf(w)。$$

- 新闻报道 d 和跟踪器 q 的相似度值采用加权和计算方法计算，其中 q_k 和 dk 分别表示各自向量中第 k 个词的权重。

$$\text{sim} (q, d) = \frac{\sum_{k=1}^N q_k \bullet d_k}{\sum_{k=1}^N q_k}$$



基于查询的跟踪技术

- if (报道d和话题的最新相关报道相距甚远)
- { 判定d为不相关报道。Stop。 }
- else if ($\text{sim}(q, d) > t1$)
- { 判定d为相关报道;
- if($\text{sim}(q, d) > t2$)
- { **重构**跟踪器q以吸收该话题重要的新特征; }
- } else { 判定d为不相关报道。 }



基于单元语言模型的跟踪技术

➤ 语言概率模型的概率比值

$$LR(D | T) = \frac{P(D | T)}{P(D)} = \frac{P(D | T)}{P(D | B)}$$

- ❖ $P(D|T)$ 表示新闻报道 D 在话题模型 T 中的条件概率，
$$P(D | T) = \prod_{w \in D} P(w | T)$$
- ❖ $P(D)$ 则是新闻报道 D 的先验概率，通常用其在背景新闻报道集合中的条件概率 $P(D|B)$ 表示， B 为背景新闻报道集合。
 - $P(D)$ 作为除数的目的类似倒排文档频率（IDF），用来降低常用词对最终概率值的影响。
- ❖ $LR(D|T)$ 的值越大，表明新闻报道 D 属于话题 T 的可能性越大。



基于分类算法的跟踪技术

- kNN算法是基于实例的分类算法
- 不要求有大量的正例训练集，不需要对词、文档有很多的先验知识。
- 由k个最相邻的数据投票决定该数据的类别归属

$$S1(\text{YES} | \vec{x}) = \sum_{\vec{d} \in P(x,k)} \cos(\vec{d}, \vec{x}) - \sum_{\vec{d} \in N(x,k)} \cos(\vec{d}, \vec{x})$$

- ❖ X表示当前新闻报道的文档向量，
- ❖ P(x,k)表示k个近邻中相关报道集合，
- ❖ N(x,k)表示k个近邻中不相关集合

$$S2(\text{YES} | \vec{x}) = \frac{1}{k1} \sum_{\vec{d} \in P(x,k1)} \cos(\vec{d}, \vec{x}) - \frac{1}{k2} \sum_{\vec{d} \in N(x,k2)} \cos(\vec{d}, \vec{x})$$



TDT评测

TDT评测



- <http://www.nist.gov/speech/tests/tdt/index.htm>
- 1996(1997)年美国国防部高级研究规划署 (DARPA) 和国家标准技术局(NIST)发起 Sponsor: DARPA; Evaluation: NIST
- **Purpose:** To develop technologies for retrieval and automatic organization of **Broadcast news** and **News wire stories** and to evaluate the performance.
- 应用：用来监控各种语言信息源，在新话题出现时发出警告，在信息安全、金融证券、行业调研等领域都有广阔的应用研究前景。

TDT评测



- Automatic Transcription: Dragon
- 语料库由语言数据联盟（LDC）提供
<http://www.ldc.upenn.edu/Projects/TDT/>
 - ❖ TDT Pilot Study -- 1997
 - ❖ TDT 2 – 1998 TDT 3 -- 1999
 - ❖ TDT 4 – 2002 TDT 5 – 2004
- TDT2 text corpus:
 - ❖ free to LDC membership; US\$500 to non-members
- 最初只针对英语语种的信息，后来加入了中文和阿拉伯语

TDT任务



- 最初只包括三个任务
 - ❖ 信息流分割 (The Story Segmentation)
 - ❖ 话题检测 (The Topic Detection)
 - ❖ 话题跟踪 (The Topic Tracking)
- 增加
 - ❖ 首篇报道检测 (The First-Story Detection)
 - ❖ 报道相关性检测 (The Link Detection)
 - ❖ Supervised Adaptive Tracking
 - ❖ 层次话题检测 (Hierarchical Topic Detection)

TDT2 — Corpus



- TDT Pilot Study (1997)
- TDT2 (1998)
 - ❖ 6 months stories, 1998.1-1998.6
 - ❖ **Text** sources: New York Times, Associated Press
 - ❖ **Video** broadcast sources: CNN, ABC
 - ❖ **Radio** broadcast sources: Voice of America, Public Radio International
 - ❖ 60,000 stories, approx. 35 million words

TDT3 — Corpus



- TDT3 (1999)
 - ❖ 3 months of news stories, 1998.10-1998.12
 - ❖ A continuation and extension of TDT2
 - ❖ In addition to TDT2, NBC and MSNBC TV broadcasts
 - ❖ 37,000 English stories
 - ❖ TDT2 and TDT3 are labeled with 120 topics, Approx. 15,000 English stories belong to one of these topics, the other are unlabeled
 - ❖ Language type: English and Mandarin(汉语)

TDT4 — Corpus



➤ TDT4 (2002)

- ❖ 4 months of news stories, 2000.10-2001.1
- ❖ Approx. 28,000 English stories,
 - Use TDT2 to **training** algorithms,
 - TDT3 to **test** algorithms duration the development period,
 - TDT4 to **evaluation** the performance of algorithms independently
- ❖ Language: English, Mandarin **and Arabic**

TDT5 — Corpus



➤ TDT5 (2004)

- ❖ April 1, 2003 to September 31, 2003
- ❖ Newswire Sources: 6 Arabic 7 English 4 Mandarin
- ❖ Broadcast News Sources: NONE
- ❖ Story Counts: 407503 news, 0 non-news
- ❖ Annotated topics: 250
- ❖ Average topic size: 40 stories



Overview of the TDT 2004 Evaluation and Results (December 2-3, 2004)

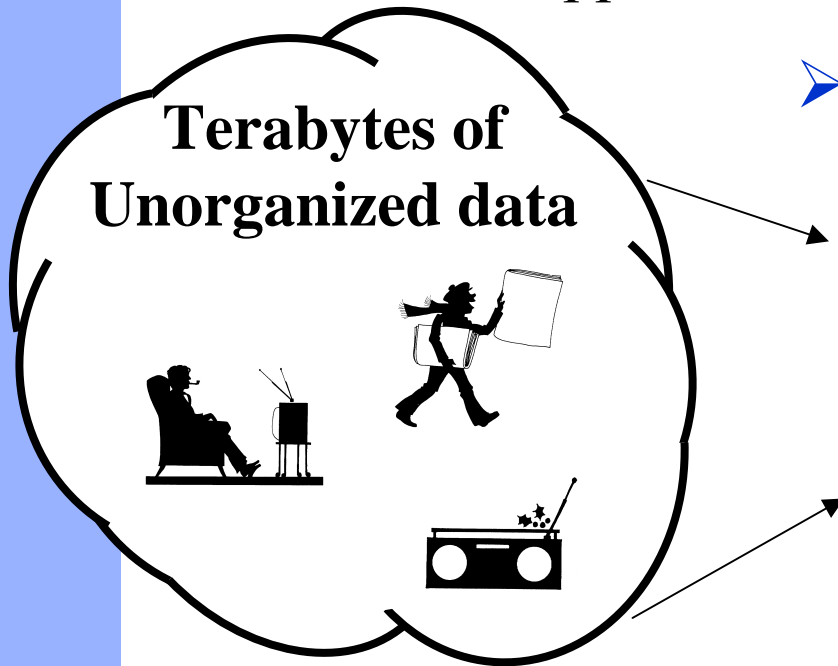
TDT 2004



“Applications for organizing text”

➤ 5 TDT Applications

- ❖ Story Segmentation*
- ❖ Topic Tracking
- ❖ Topic Detection
- ❖ First Story Detection
- ❖ Link Detection



* Not evaluated in 2004

TDT's Research Domain



- Technology challenge
 - ❖ Develop applications that organize and locate relevant stories from a **continuous** feed of news stories
- Research driven by evaluation tasks
- Composite applications built from
 - ❖ Document Retrieval
 - ❖ Speech-to-Text (STT) – **not** included this year
 - ❖ Story Segmentation – **not** included this year

Evaluation Corpus



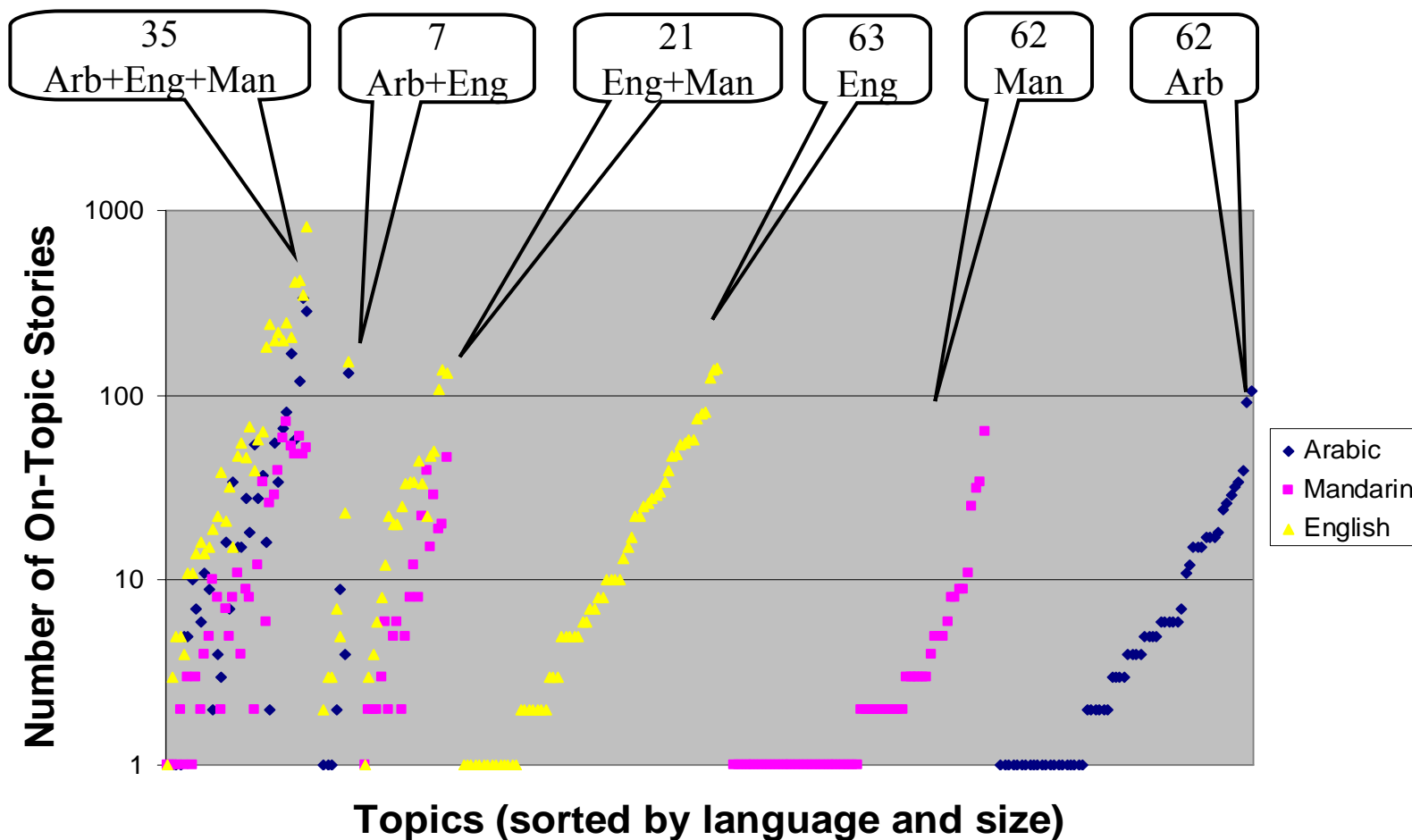
- Same languages as 2003
- Summary of differences
 - ❖ New time period
 - ❖ No broadcast news
 - No non-news stories
 - ❖ 4.5 times more stories
 - ❖ 3.1 times more topics
 - ❖ Topics have $\frac{1}{2}$ as many on-topic stories

Evaluation Corpus



	TDT4 (2003's corpus)	TDT5 (2004's corpus)
Collection Dates	October 1, 2000 to January 31, 2001	April 1, 2003 to September 31, 2003
Newswire Sources	3 Arabic 2 English 2 Mandarin	6 Arabic 7 English 4 Mandarin
Broadcast News Sources	2 Arabic 5 English 5 Mandarin	NONE
Story Counts	90735 news, 7513 non-news stories	407503 news, 0 non-news
Annotated topics	80	250
Average topic size	79 stories	40 stories

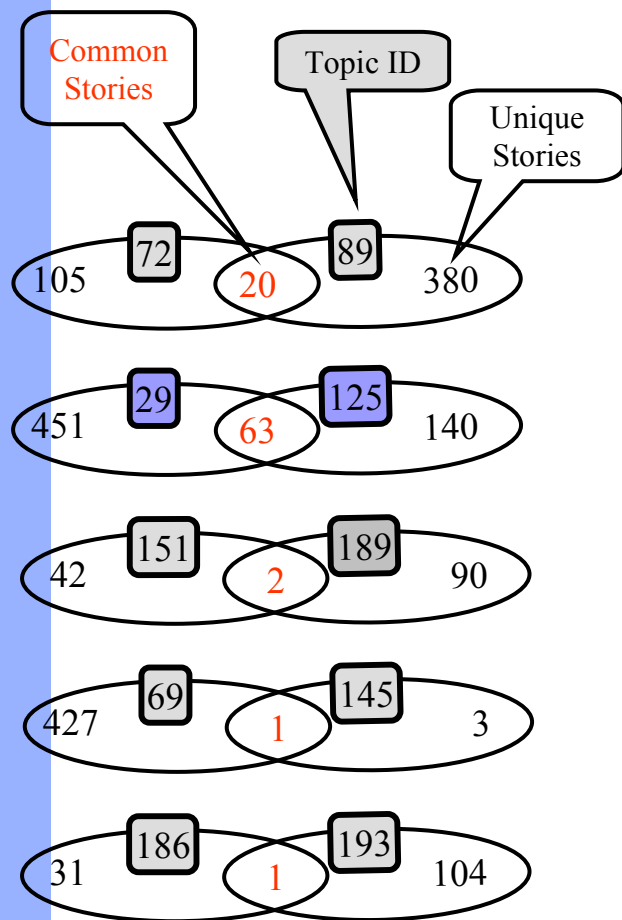
Topic Size Distribution



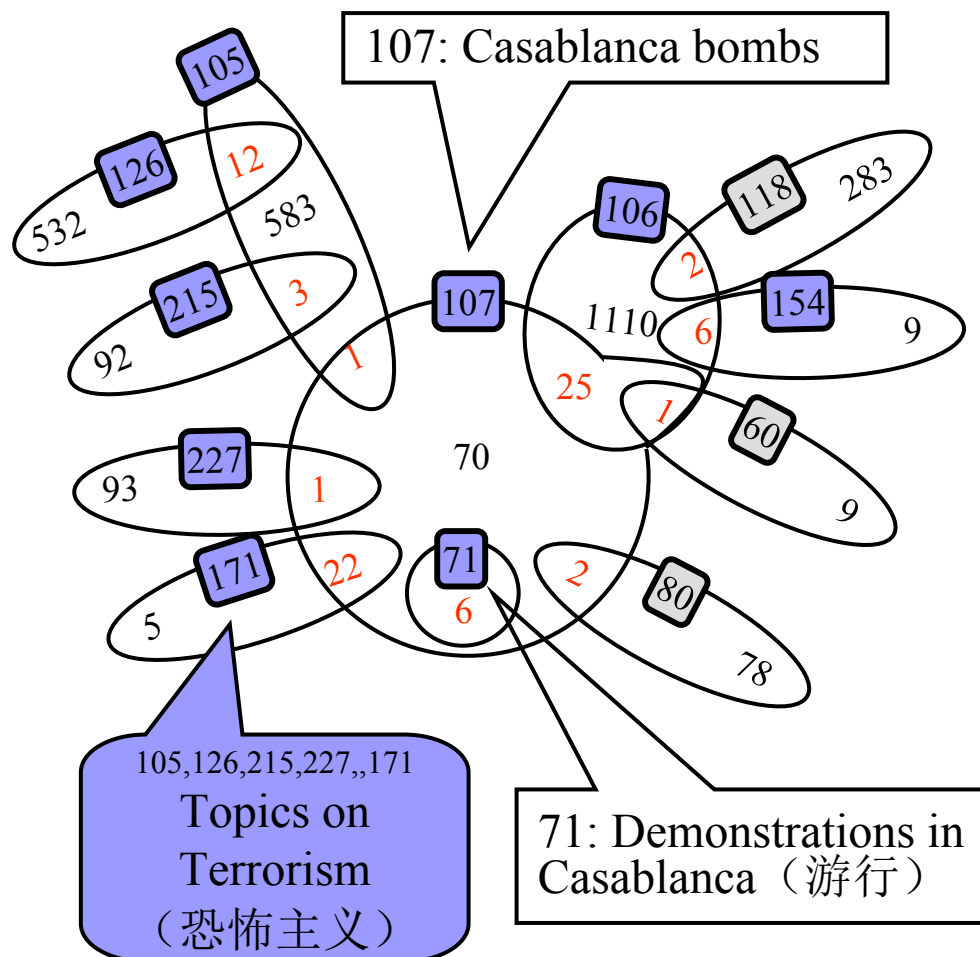


Topic overlap

Single **Overlap** Topics



Multiply Overlap Topics



Topic labels



Single Overlap Topics

- 72** Court indicts Liberian President
- 89** Liberian former president arrives in exile
- 29** Swedish Foreign Minister killed
- 125** Sweden rejects the Euro
- 151** Egyptian delegation in Gaza
- 189** Palestinian public uprising suspended for three months
- 69** Earthquake in Algeria
- 145** Visit of Morocco Minister of Foreign Affairs to Algeria
- 186** Press conference between Lebanon and US foreign ministers
- 193** Colin Powell Plans to visit Middle East and Europe

Multiply Overlap Topics

- 105** UN official killed in attack
- 126** British soldiers attacked in Basra
- 215** Jerusalem: Bus suicide bombing
- 227** Bin Laden Videotape
- 171** Morocco: death sentences for bombing suspects
- 107** Casablanca bombs
- 71** Demonstrations in Casablanca
- 106** Bombing in Riyadh, Saudi Arabia
- 118** World Economic Forum in Jordan
- 154** Saudi suicide bomber dies in shootout
- 60** Saudi King has eye surgery
- 80** Spanish Elections

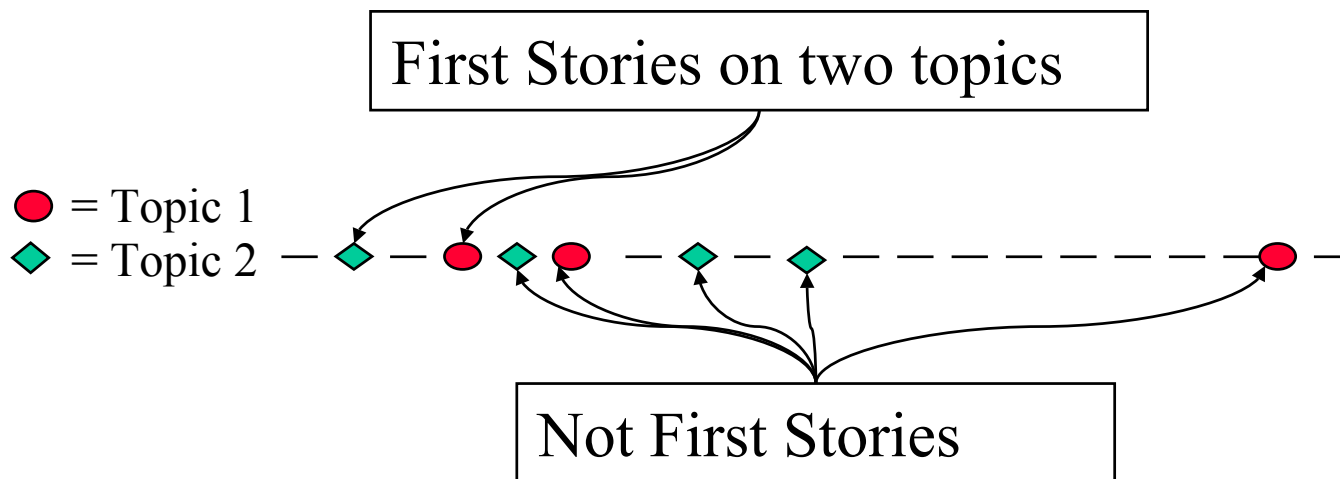


New Event Detection Task



System Goal:

- ❖ To detect the first story that discusses each topic



TDT Evaluation Methodology



- Tasks are modeled as detection tasks
 - ❖ Systems are presented with many trials and must answer the question:

“Is this example a **target** trial?”
 - ❖ Systems respond:
 - YES this is a target, or NO this is not
 - Each decision includes a likelihood score indicating the system’s **confidence** in the decision
- System performance measured by linearly combining the system’s **missed detection rate** and **false alarm rate**

Detection Evaluation Methodology



➤ Performance is measured in terms of **Detection Cost**

$$❖ C_{\text{Det}} = C_{\text{Miss}} * \mathbf{P_{\text{Miss}}} * P_{\text{target}} + C_{\text{FA}} * \mathbf{P_{\text{FA}}} * (1 - P_{\text{target}})$$

❖ Constants:

- $C_{\text{Miss}} = 1$ and $C_{\text{FA}} = 0.1$ are preset costs
- $P_{\text{target}} = 0.02$ is the **a priori probability** of a target

❖ System performance estimates

- $\mathbf{P_{\text{Miss}}}$ and $\mathbf{P_{\text{FA}}}$
- Miss: $\text{miss} = c/(a+c)$ False alarm (f): $f = b/(b+d)$

❖ **Normalized** Detection Cost generally lies between 0 and 1:

- $(C_{\text{Det}})_{\text{Norm}} = C_{\text{Det}} / \min \{ C_{\text{Miss}} * P_{\text{target}}, C_{\text{FA}} * (1 - P_{\text{target}}) \}$

Detection Evaluation Methodology

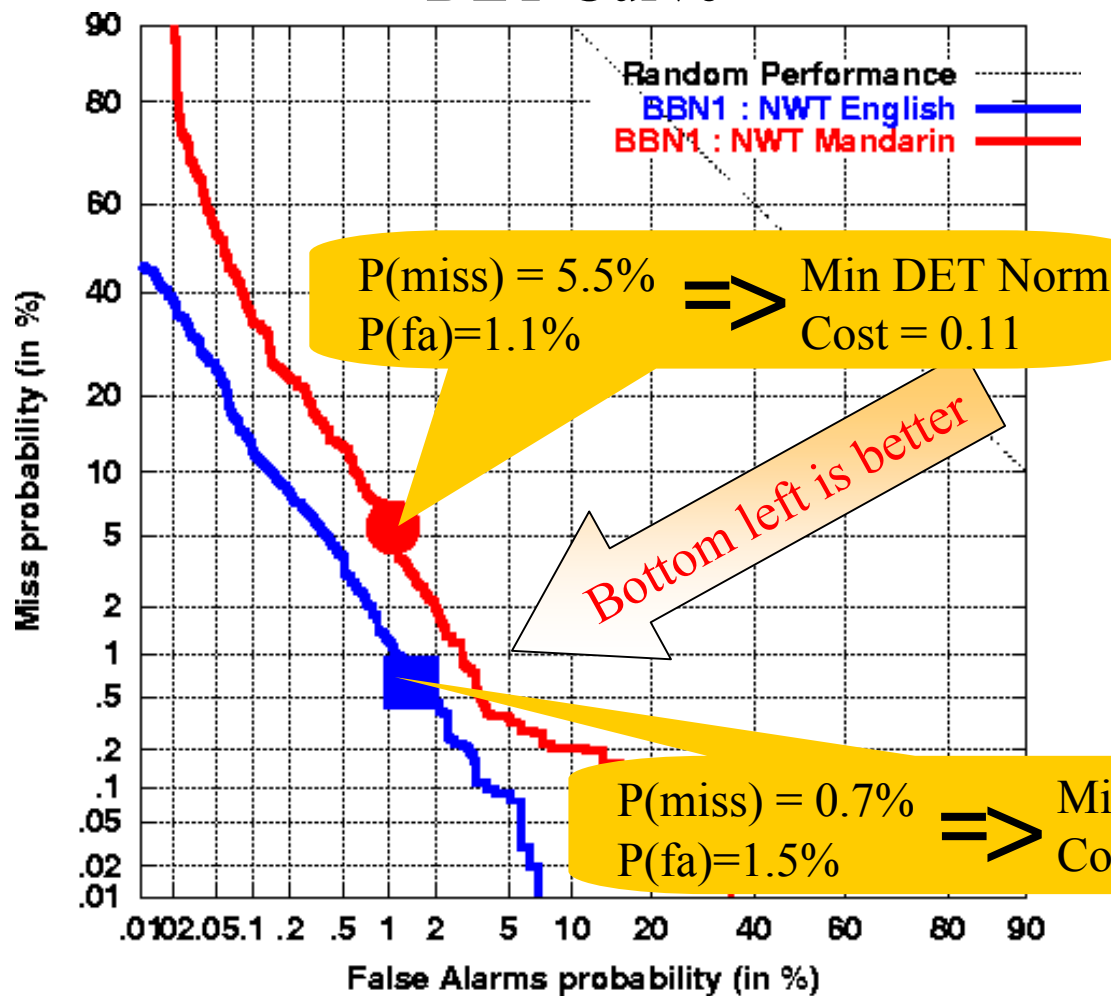


- **Detection Error Tradeoff (DET) curves** graphically depict the performance tradeoff between P_{Miss} and P_{FA}
 - ❖ Makes use of likelihood scores attached to the YES/NO decisions
- ✧ Two important scores per system
 - ❖ Actual Normalized **Detection Cost**
 - Based on the YES/NO decision threshold
 - ❖ Minimum Normalized **DET point**
 - Based on the DET curve: Minimum score with proper threshold

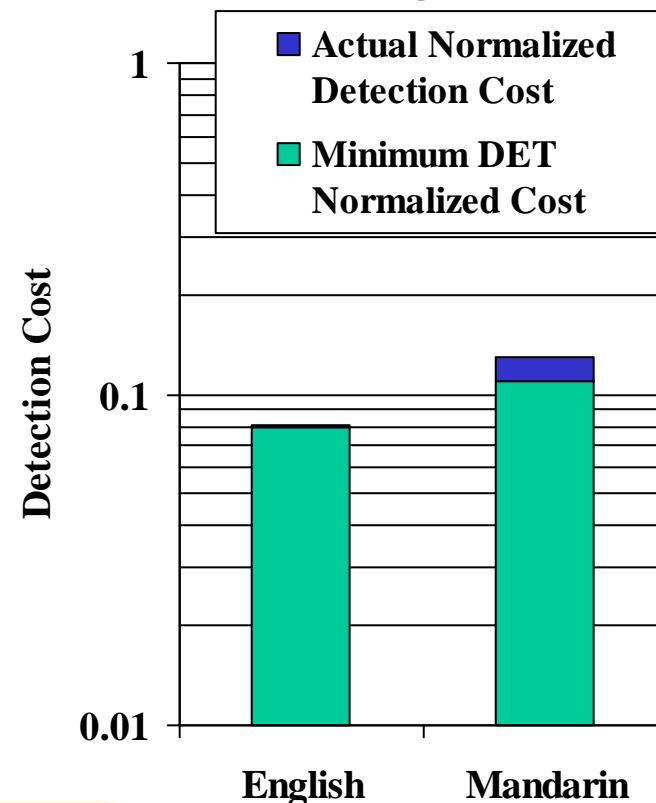
Performance Measures Example



DET Curve

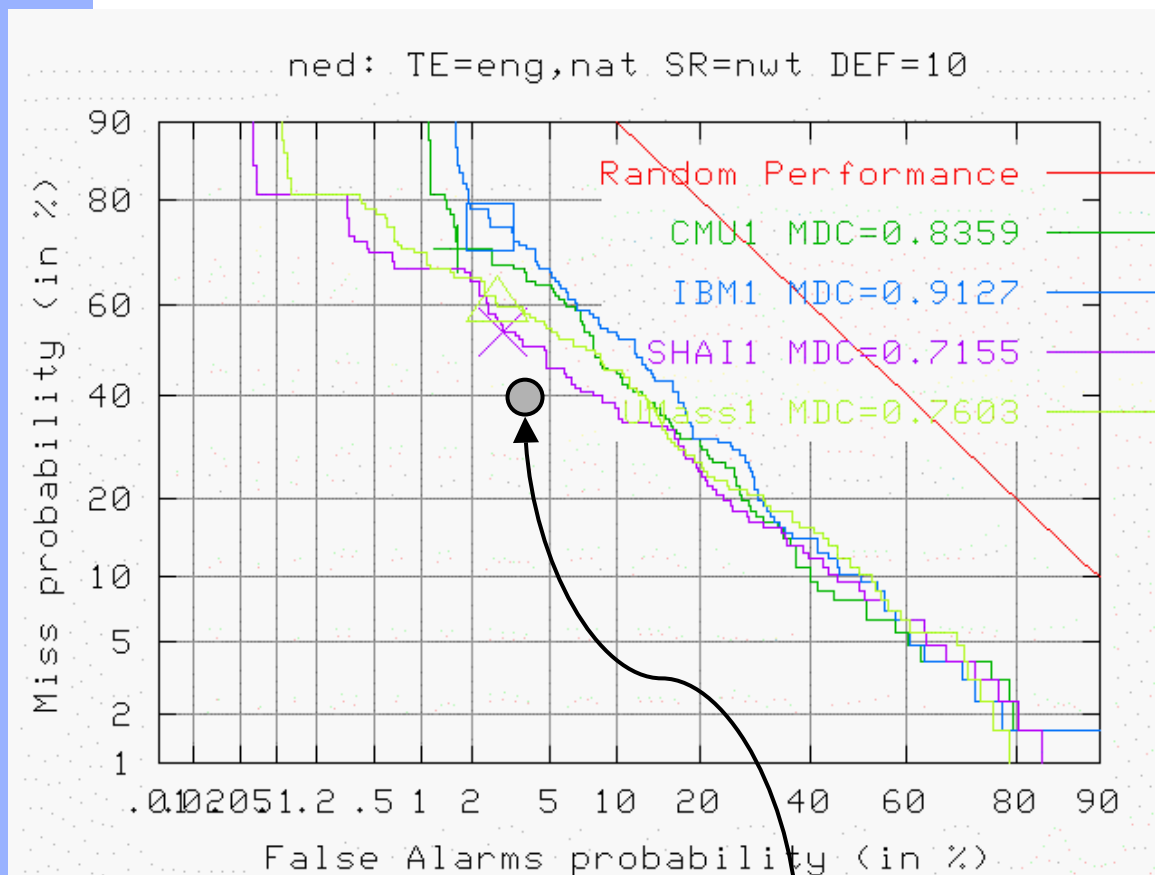


Bar Chart

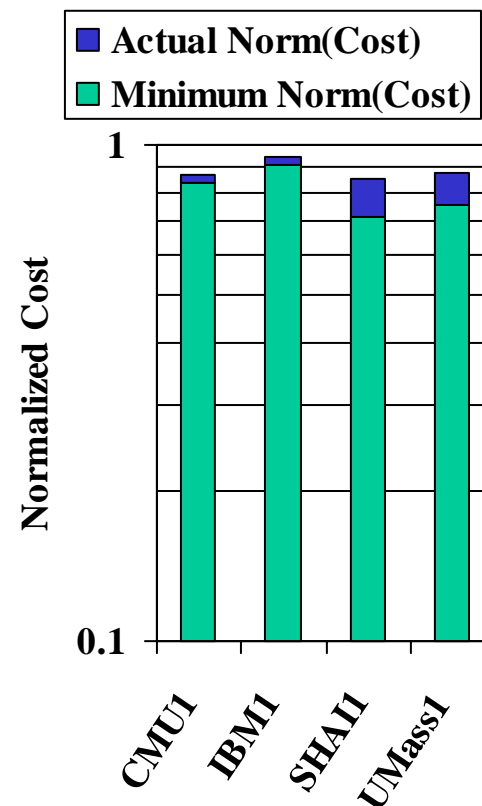


Primary New Event Detection Results

Newsire, English Texts



2003's best score



New Event Detection Performance History



<i>year</i>	<i>condition</i>	<i>site</i>	<i>score</i>
1999	SR=nwt+bnasr TE=eng,nat boundary DEF=10	UMass1	.8110
2000	SR=nwt+bnasr TE=eng,nat noboundary DEF=10	UMass1	.7581
2001	“ “	UMass1	.7729
2002	SR=nwt+bnasr TE=eng,nat boundary DEF=10	CMU1	.4449
2003	“”	CMU1	.5971*
2004	SR=nwt TE=eng,nat DEF=10	UMass2	.8387

* 0.4283 on 2002 Topics₈₈

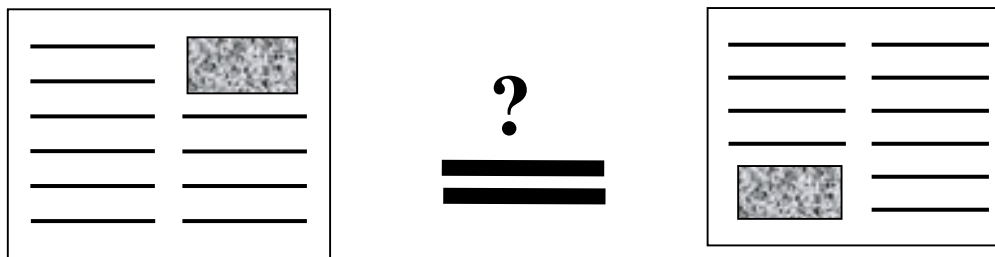
TDT Link Detection Task



System Goal:

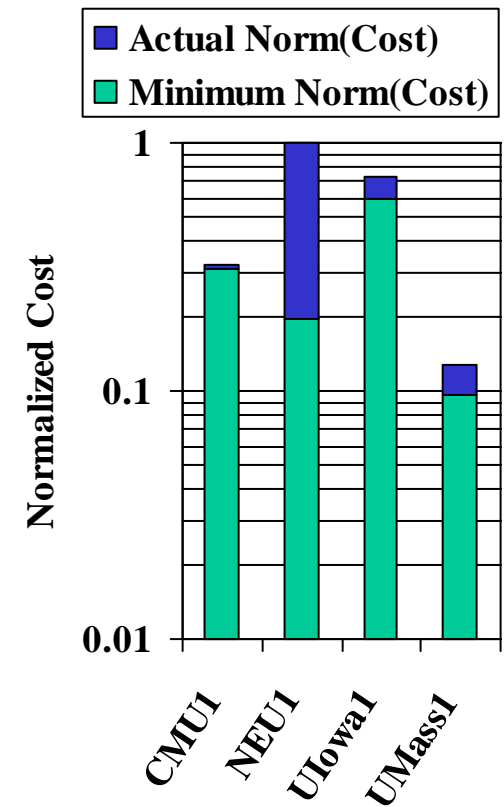
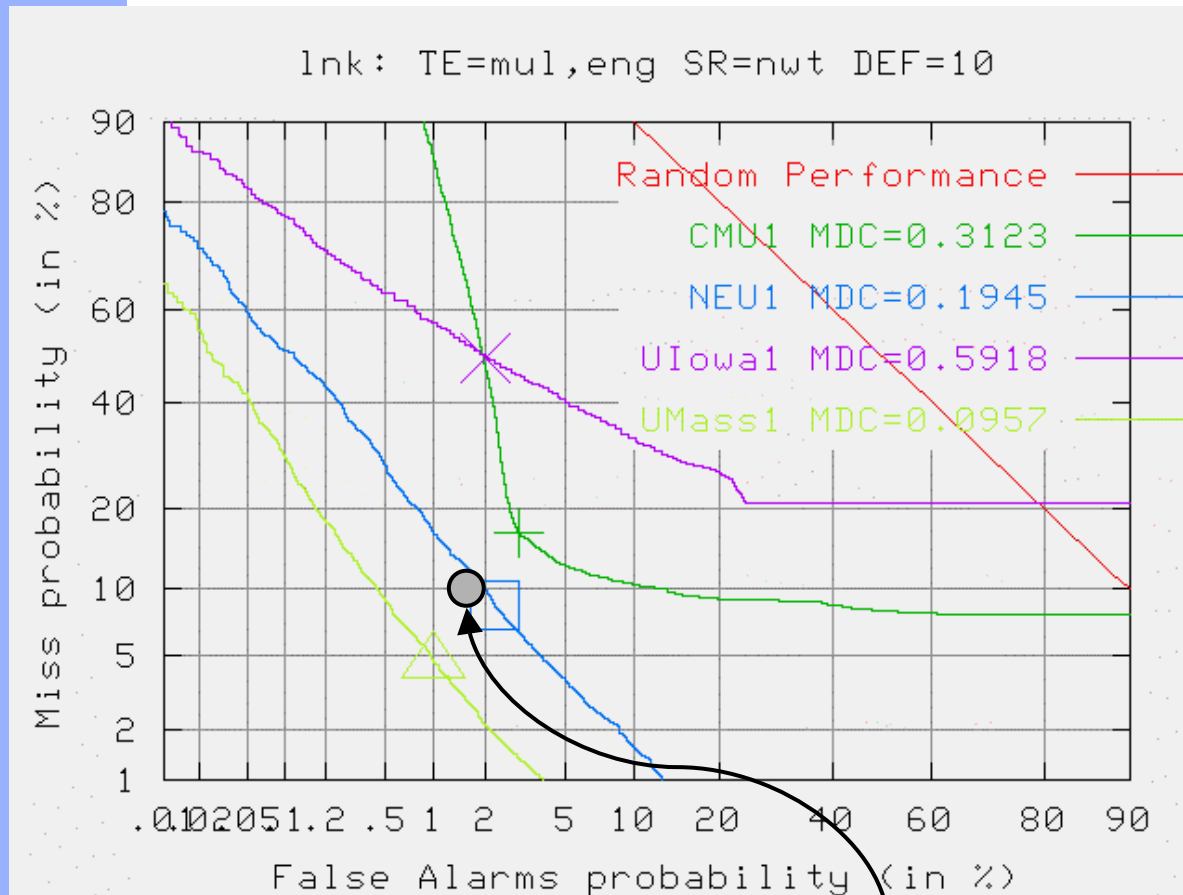
- ❖ To detect whether a pair of stories discuss the **same** topic.

(Can be thought of as a “**primitive operator**” to build a variety of applications)



Primary Link Detection Results

Newsire, Multilingual links, 10-file deferral period



Scores are better than 2003!

Link Detection Performance History



<i>year</i>	<i>condition</i>	<i>site</i>	<i>score</i>
1999	SR=nwt+bnasr TE=eng,nat DEF=10	CMU1	1.0943
2000	SR=nwt+bnasr TE=eng+man,eng boundary DEF=10	UMass1	.3134
2001	“ “	CMU1	.2421
2002	SR=nwt+bnasr TE=eng+man+arb, eng boundary DEF=10	PARC1	.1947
2003	SR=nwt+bnasr TE=eng+man+arb, eng boundary DEF=10	UMass01	.1839*
2004	SR=NWT TE=eng+man+arb DEF=10	CMU6	0.1047

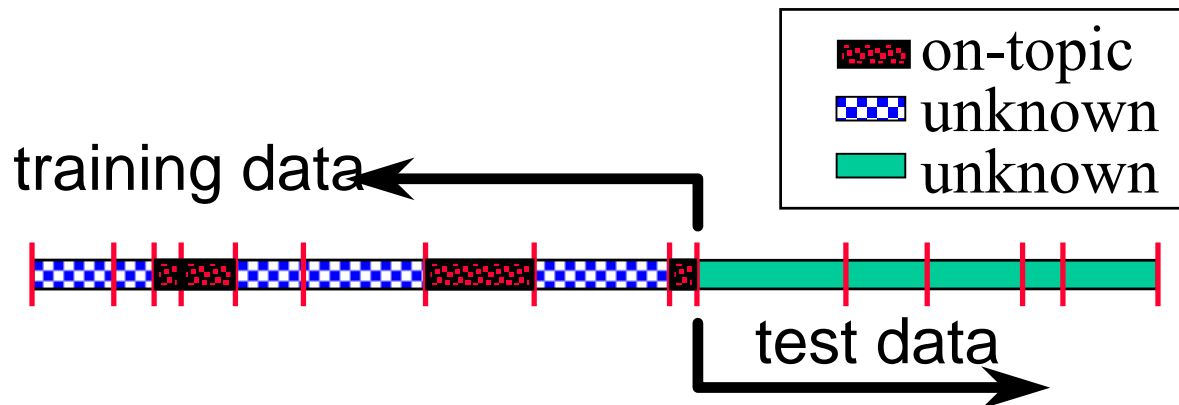
* 0.1798 on 2002 Topics₉₁

Topic Tracking Task



➤ System Goal:

- ❖ To detect stories that discuss the target topic, in multiple source streams
 - Supervised Training
 - Given N_t samples stories that discuss a given target topic
 - Testing
 - Find all subsequent stories that discuss the target topic

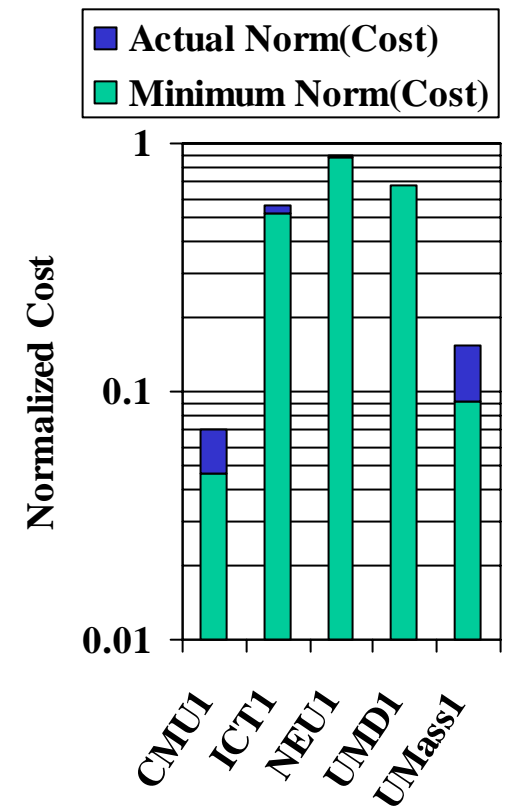
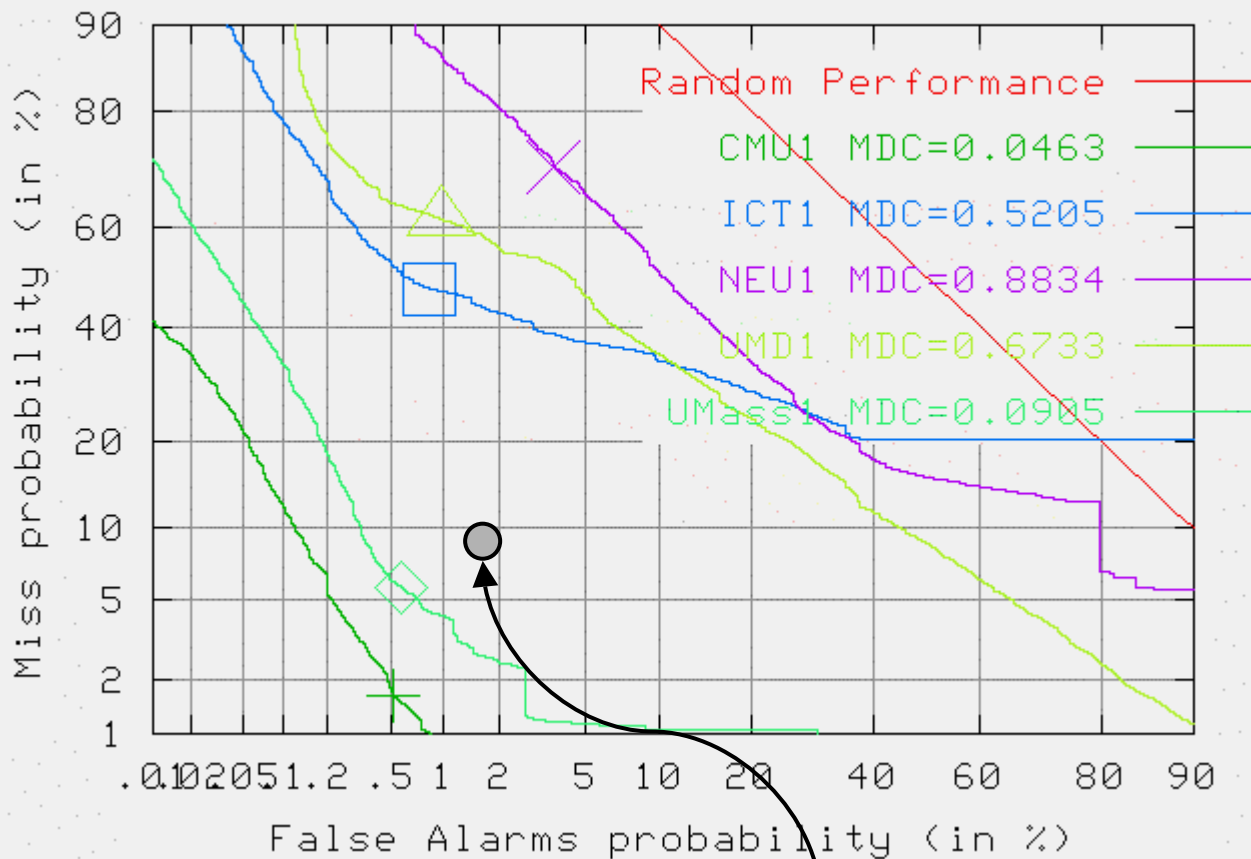


Primary Tracking Results

Newsire, Multilingual Texts, 1 English Training Story



tracking: TE=mul,eng SR=nwt TR=eng Nt=1



Tracking Performance History



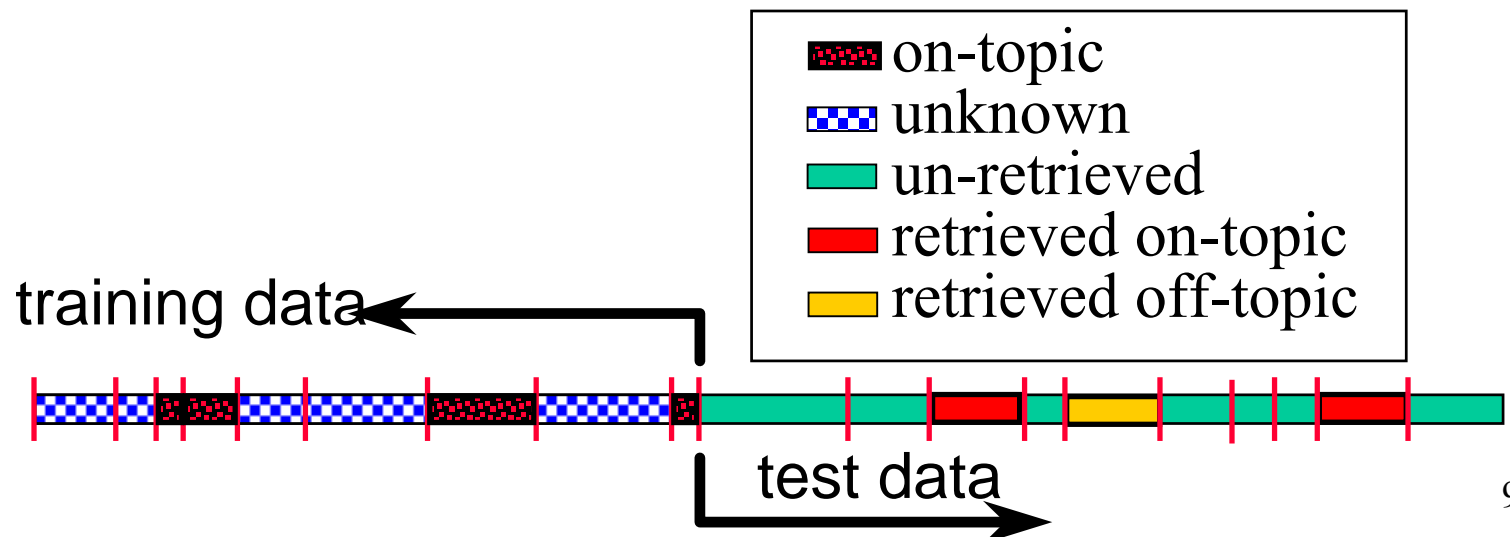
<i>year</i>	<i>condition</i>	<i>site</i>	<i>score</i>
1999	SR=nwt+bnasr TR=eng TE=eng+man,eng boundary NT=4	BBN1	.0922
2000	SR=nwt+bnman TR=eng TE=eng+man,eng boundary NT=1_Nn=0	IBM1	.1248
2001	“ “	LIMSI1	.1213
2002	SR=nwt+bnman TR=eng TE=eng+man+arb, eng boundary Nt=1 Nn=0	UMass1	.1647
2003	SR=nwt+bnman TR=eng TE=eng+man+arb, eng boundary Nt=1 Nn=0	UMass1	.1949*
2004	SR=nwt TR=eng TE=eng+man+arb Nt=1	CMU2	.0599

* 0.1618 on 2002 Topics

Supervised Adaptive Tracking Task



- Variation of Topic Tracking system goal:
 - ❖ To detect stories that discuss the target topic **when a human provides feedback to the system**
 - System receives human judgment (on or off-topic) for every retrieved story
 - ❖ Same task as TREC 2002 Adaptive Filtering



Supervised Adaptive Tracking Metrics



- Normalized Detection Cost
 - ❖ Same measure as for basic Tracking task
- Linear Utility Measure
 - ❖ As defined for TREC 2002 Filtering Track (Robertson & Soboroff)
 - ❖ Measures value of the stories sent to the user:
 - Credit for relevant stories, debit for non-relevant stories
 - Equivalent to thresholding based on estimated probability of relevance
 - ❖ No penalty for missing relevant stories (i.e. all precision, no recall)
 - ❖ Implication: Challenge is to beat the “do-nothing” baseline (i.e. a system that rejects all stories)



➤ Linear Utility Measure Computation:

❖ Basic formula: $U = W_{\text{rel}} \times R - NR$

- R = number of **relevant** stories retrieved
- NR = number of **non-relevant** stories retrieved
- W_{rel} = relative **weight** of relevant vs non-relevant (set to 10, by analogy with C_{Miss} vs. C_{FA} weights for C_{Det})

❖ Normalization across topics:

- Divide by **maximum** possible utility score for each topic



➤ Linear Utility Measure Computation:

❖ Scaling across topics:

- Define arbitrary minimum possible score, to avoid having average dominated by a few topics with huge NR counts
- Corresponds to application scenario in which user stops looking at stories when system **exceeds some tolerable false alarm rate**

❖ Scaled, normalized value:

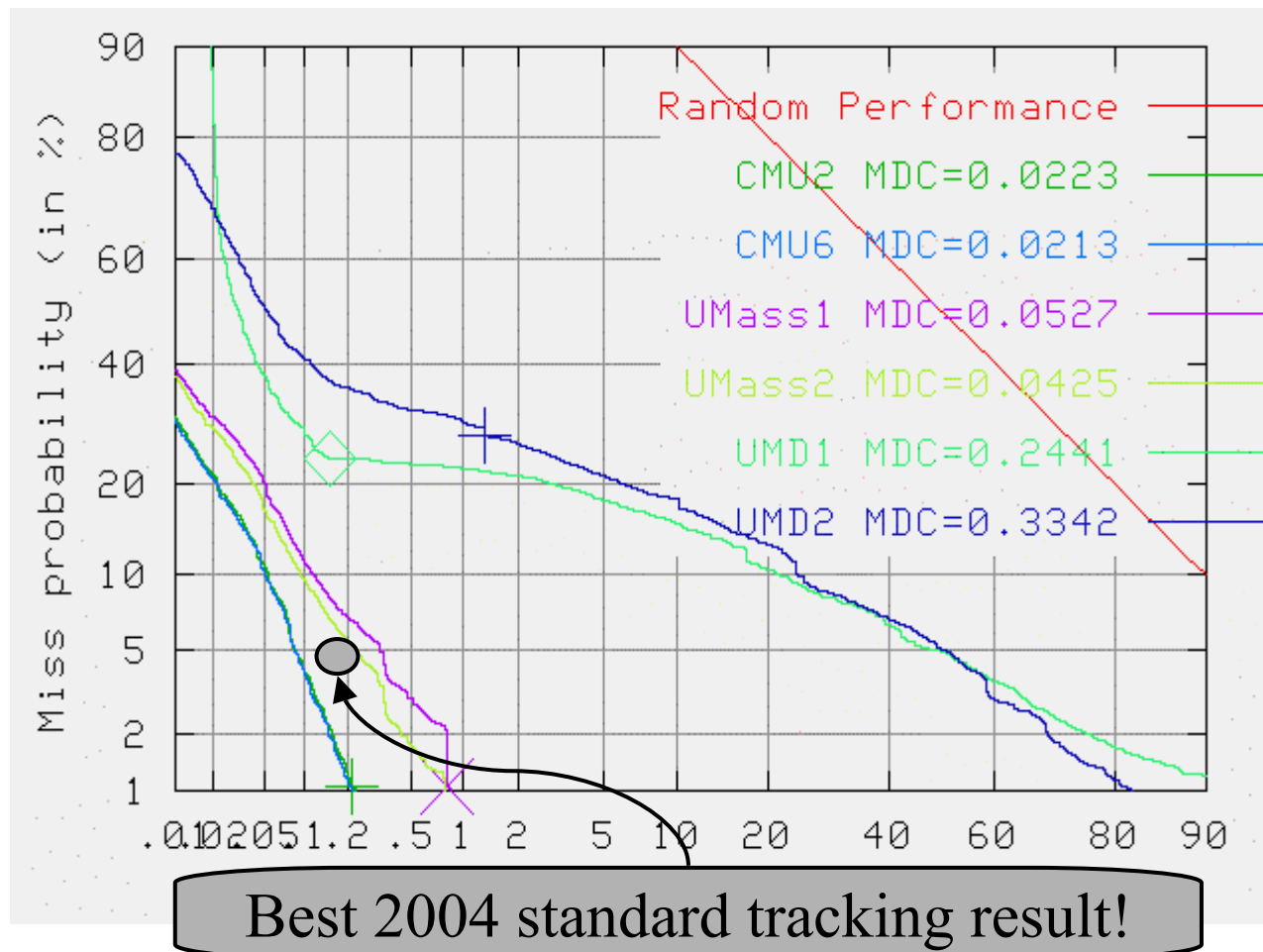
$$U_{\text{scale}} = [\max(U_{\text{norm}}, U_{\text{min}})] / [1 - U_{\text{min}}]$$

Supervised Adaptive Tracking

Best Two Submissions per Site



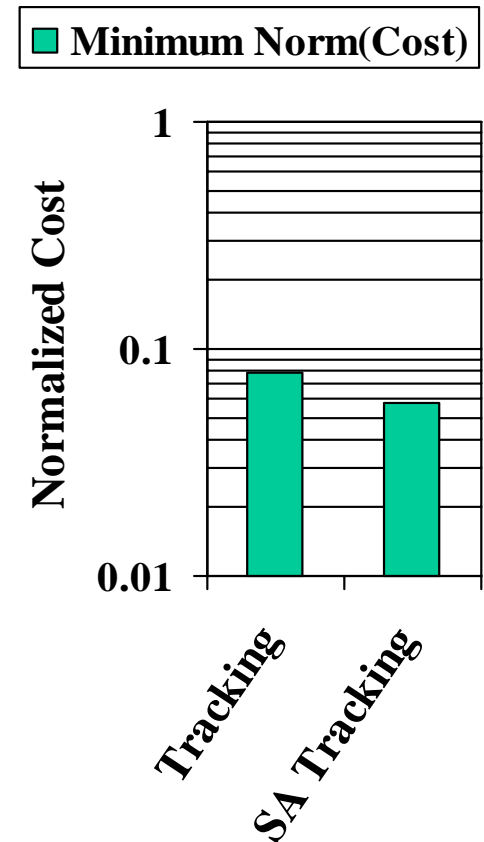
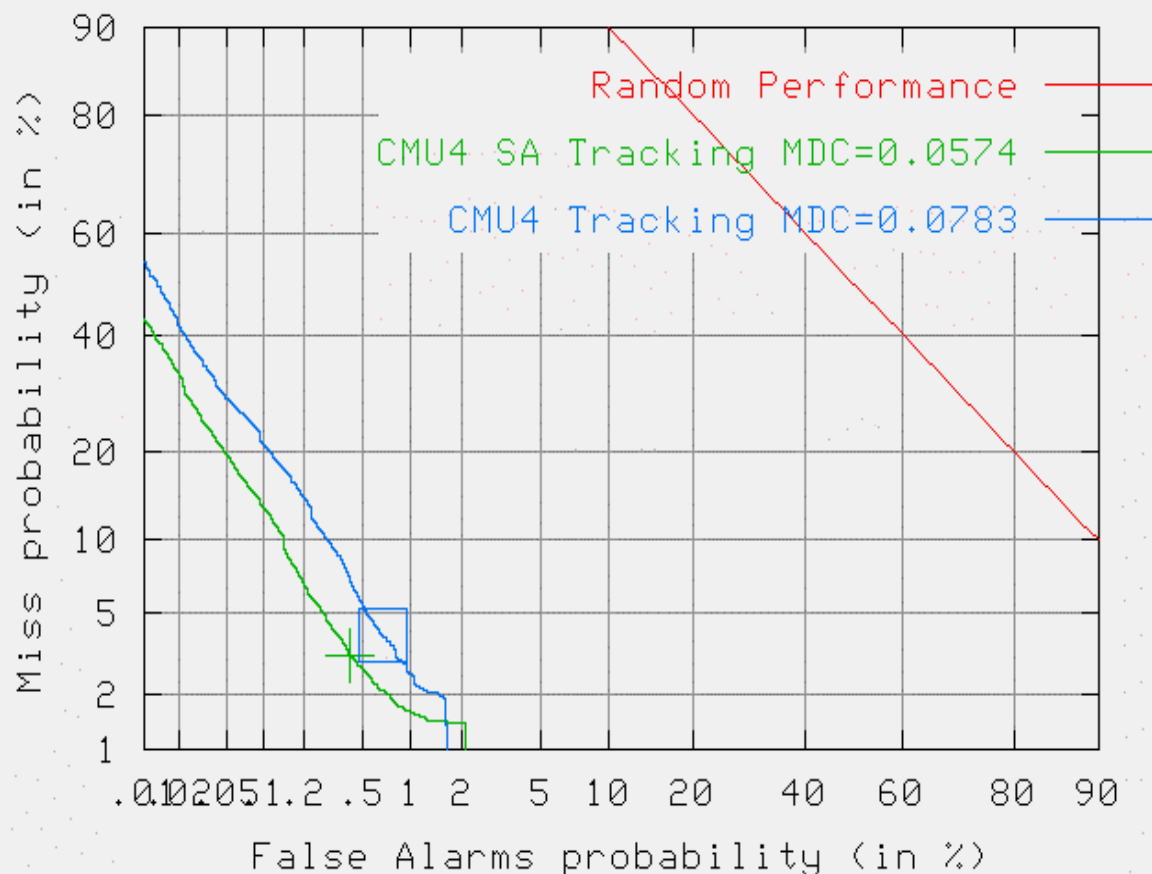
Newswire, Multilingual Texts, 1 English Training Story



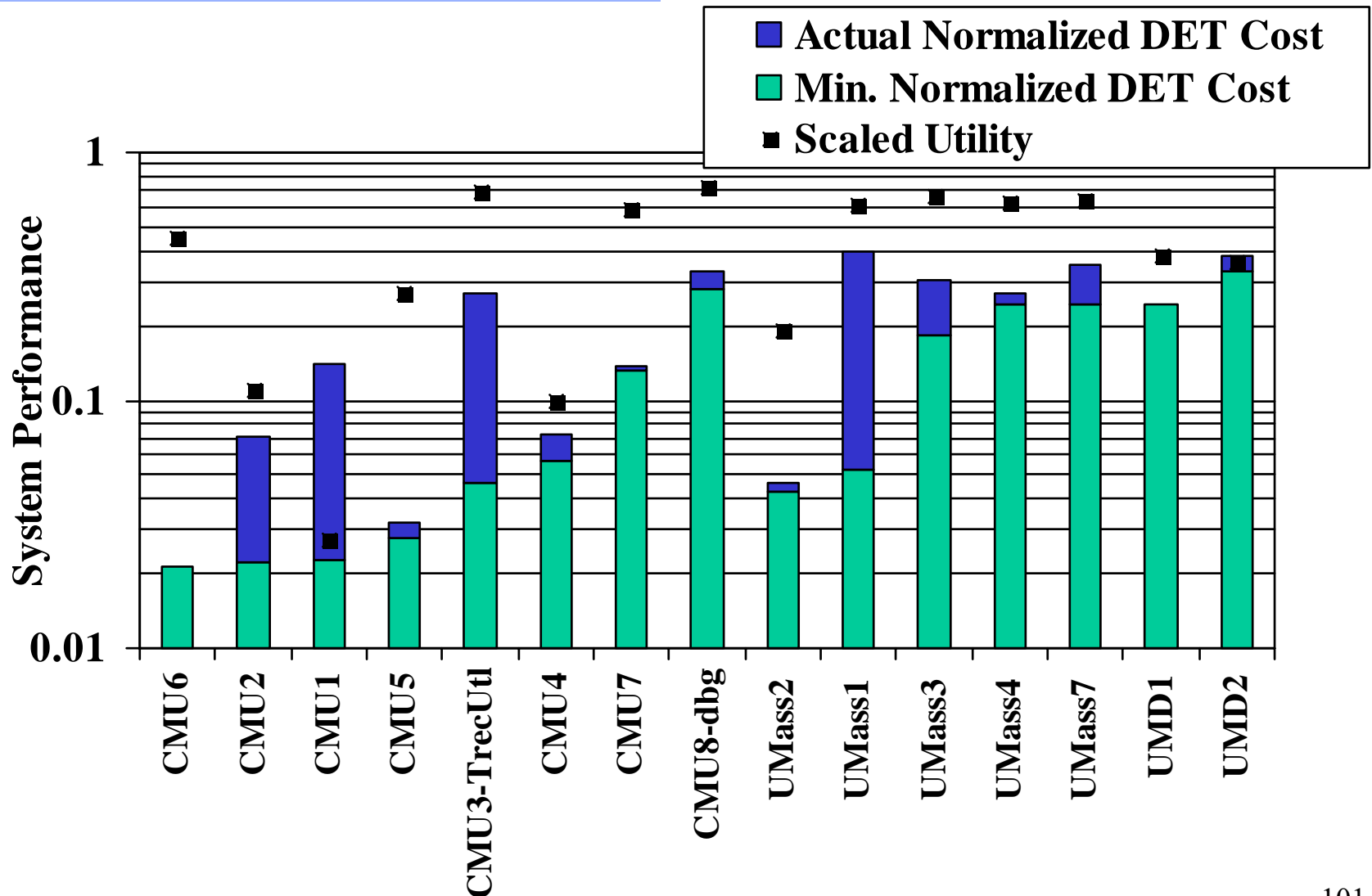
Effect of Supervised Adaptation



- CMU4 is a simple cosine similarity tracker
 - ❖ **Contrastive** run submitted **without supervised** adaptation



Supervised Adaptive Tracking

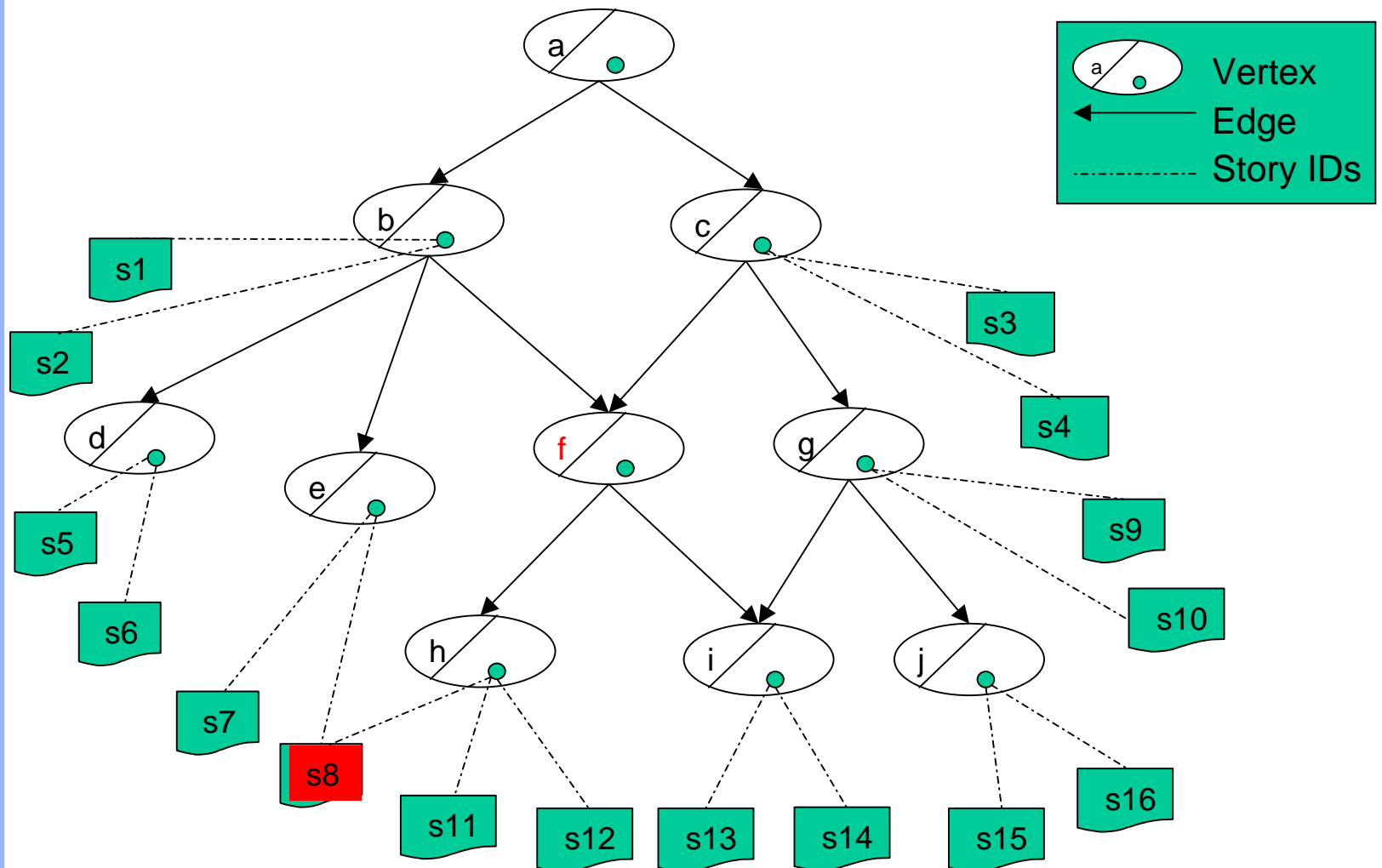


Hierarchical Topic Detection



- System goal:
 - ❖ To detect topics in terms of the (clusters of) stories that discuss them
- Problems with past Topic Detection evaluations:
 - ❖ Topics are at **different levels** of granularity, yet systems had to choose single operating point for creating a new cluster
 - ❖ Stories may pertain to **multiple topics**, yet systems had to assign each to only one cluster

Topic Hierarchy Solves Problems



Topic Hierarchy Solves Problems



➤ System operation:

- ❖ **Unsupervised** topic training - no topic instances as input
- ❖ Assign each story to one or **more** clusters
- ❖ Clusters may **overlap** or include other clusters
- ❖ Clusters must be organized as **directed acyclic graph (DAG)** with single root
- ❖ Treated as retrospective search

Topic Hierarchy Solves Problems



- Semantics of **topic hierarchy**:
 - ❖ Root = entire collection
 - ❖ Leaf nodes = the most specific topics
 - ❖ Intermediate nodes represent different levels of granularity
- Performance assessment:
 - ❖ Given a topic, find matching cluster with **lowest cost**

Hierarchical Topic Detection

Metric: Minimal Cost



- Minimal Cost metric selected based on study at U Mass (Allan et al.):
 - ❖ Effectively eliminates power set solution
 - ❖ Favors balance of **cluster purity** vs. number of clusters
 - ❖ Computationally tractable
 - ❖ Good behavior in U Mass experiments

Hierarchical Topic Detection

Metric: Minimal Cost



- Weighted combination of **Detection Cost (识别代价)** and **Travel Cost (遍历代价)**:

$$\text{WDET} \times (\text{Cdet}(\text{topic}, \text{bestVertex}))_{\text{Norm}} + (1 - \text{WDET}) \times \text{Ctravel}(\text{topic}, \text{bestVertex})_{\text{Norm}}$$

- ❖ Detection Cost: same as for other tasks
- ❖ Travel Cost: function of the hierarchy
- ❖ Detection Cost weighted 2× Travel Cost (**WDET = 0.66**)

Hierarchical Topic Detection

Metric: Travel Cost



- Travel Cost computation: (遍历代价)
$$\text{Ctravel}(\text{topic}, \text{vertex}) = \text{Ctravel}(\text{topic}, \text{parentOf}(\text{vertex})) + \text{CBRANCH} \times \text{NumChildren}(\text{parentOf}(\text{vertex})) + \text{CTITLE}$$
 - ❖ CBRANCH = cost per branch, for each vertex on path to best match (两个预设参数)
 - ❖ CTITLE = cost of examining each vertex
 - ❖ Relative values of CBRANCH and CTITLE determine preference for shallow, bushy hierarchy vs. deep, less bushy hierarchy
 - ❖ Evaluation values chosen to favor branching factor of 3

Hierarchical Topic Detection

Metric: Travel Cost



➤ Travel Cost **normalization**:

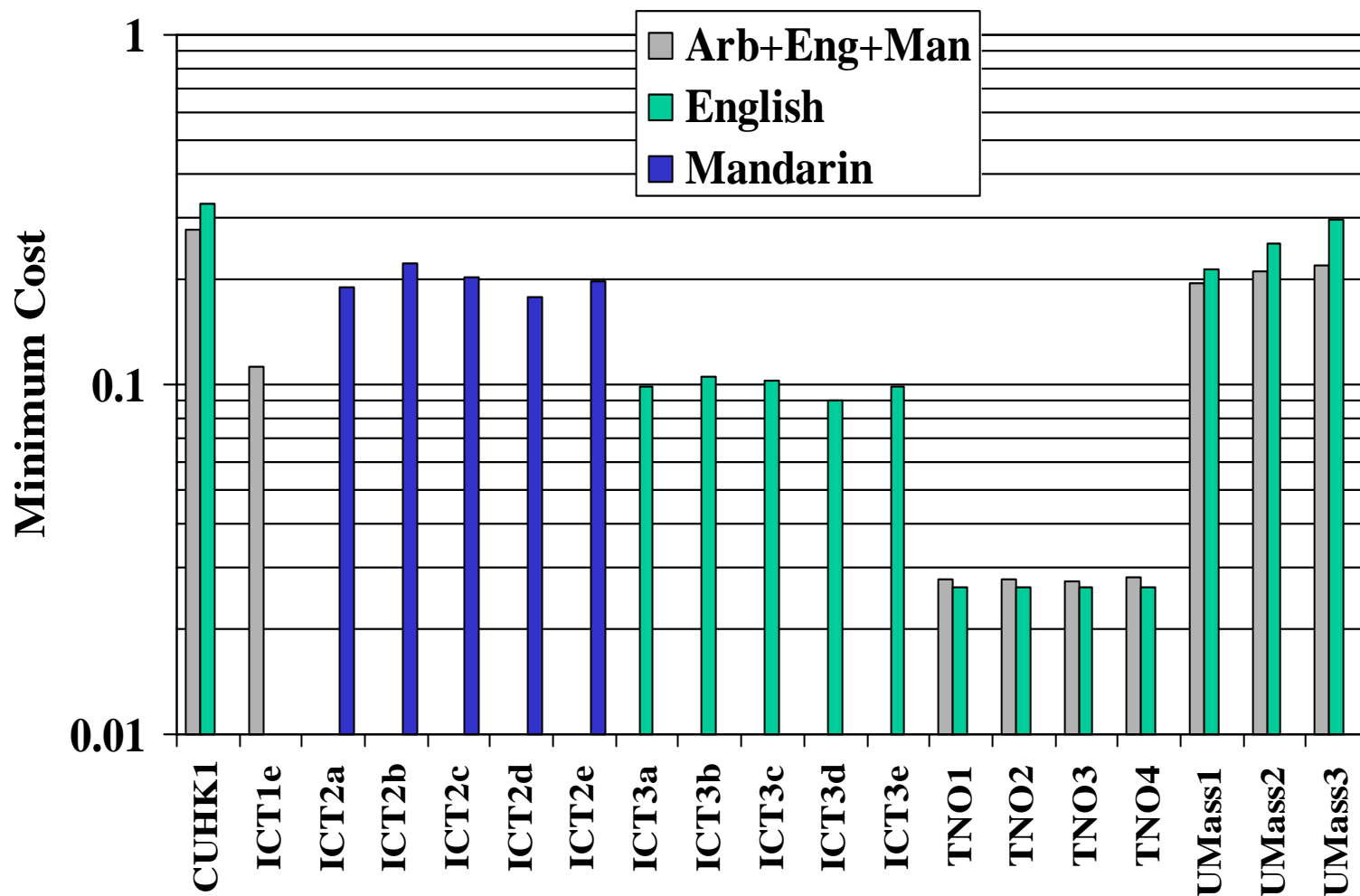
- ❖ Absolute travel cost depends on **size of corpus**, diversity of topics
- ❖ Must be normalized to combine with Detection Cost
- ❖ Normalization scheme for trial evaluation chosen to yield $C_{\text{travel}}^{\text{Norm}} = 1$ for “ignorant” hierarchy (by analogy with use of prior probability for $C_{\text{det}}^{\text{Norm}}$):

$$C_{\text{travel}}^{\text{Norm}} = \frac{C_{\text{travel}}}{(\text{CBRANCH} * \text{MAXVTS} * \text{NSTORIES} / \text{AVESPT}) + \text{CTITLE}}$$

$\text{MAXVTS} = 3$ (maximum number of vertices per story, controls **overlap**)

$\text{AVESPT} = 88$ (**average stories per topic**, computed from TDT4 multilingual data)

Hierarchical Topic Detection



Hierarchical Topic Detection Observations



- All systems structured hierarchy as a **tree** — each vertex has **one parent**
- **Travel cost** has very little effect on finding the best cluster
 - ❖ Setting WDET to 1.0 has little effect on topic mapping
- Cost parameters favor **false alarms**
 - ❖ Average mapped cluster sizes are between 1262 and 7757 stories
 - ❖ Average topic size is 40 stories

Summary



- Eleven research groups participated in five evaluation tasks
- Error rates increased for new event detection
 - ❖ Why?
- Error rates decreased for tracking
- Error rates decreased for link detection

Summary



- Dry run of **hierarchical topic detection** completed
 - ❖ Solves previous problems with topic detection task, but raises new issues
 - ❖ Questions to consider:
 - Is the specified hierarchical structure (single-root DAG) appropriate?
 - Is the minimal cost metric appropriate?
 - If so, is the normalization right?
- Dry run of **supervised adaptive tracking** completed
 - ❖ Promising results for including relevance **feedback**
 - ❖ Questions to consider:
 - Should we continue the task?
 - If so, should we continue using both metrics?

小结



- TDT相关概念
- TDT技术
- TDT评测概述
- TDT2004评测

思考题



➤ 如何检测分析互联网新闻/论坛热点

❖ 热点检测

- 内容多，时间持续，如何平衡正确性和效率？
- 话题多，如何排序？

❖ First Story Detection（首发、溯源）

- 检测时间 \neq 发布时间 \neq 网页标识时间

❖ 网站间传播关系？

❖ 话题间的演化关系？

❖ 观点分析？

❖ ? ? ? ?

任选一个作为课程上机作业



Any Question?