自动文档摘要

社交网络与舆情预测 第三讲

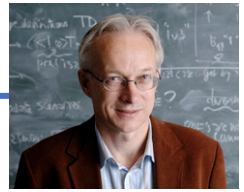
任昭春

?

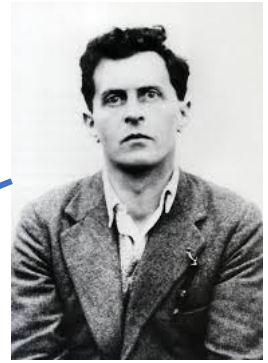
Zhaochun Ren

Maarten de Rijke

Johan F. A. K. van Benthem

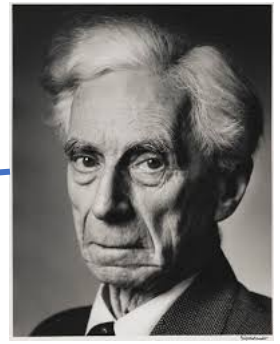


Martin Hugo Löb



Reuben Louis Goodstein

Ludwig Wittgenstein



Bertrand Arthur William Russell

....

Isaac Newton

From The Mathematics Genealogy

学术硕士：3名
专业硕士：2-3名

http://ir.sdu.edu.cn/~zhaochunren/

方向：

信息检索
- 推荐系统
- 知识图谱
- 个性化检索
- 社交计算

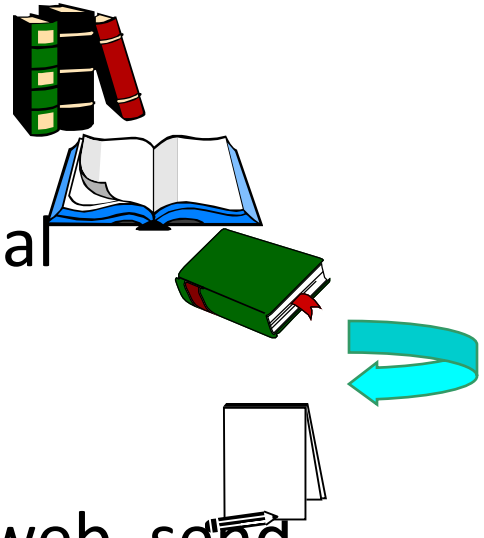
自然语言处理
- 对话生成
- 问答系统
- 自动摘要
- 阅读理解

一线互联网公司核心算法团队6个月-1年实习+推荐
海外top高校交流+推荐
Research-oriented

Outline

- 自动摘要简介
- 自动摘要方法介绍
- 最新研究进展

An exciting challenge...



...put a book on the scanner, turn the dial to ‘2 pages’, and read the result...

...download 1000 documents from the web, send them to the summarizer, and select the best ones by reading the summaries of the clusters...

...forward the Japanese email to the summarizer, select ‘1 par’, and skim the translated summary.

Headline news — informing

TIME.com | HOME | SEARCH |

TIME Daily
> News Wire
> Editor's Letter
> Comments
> News Features
> Text Only

Magazine
Community
Special Reports

LIFE Picture of the Day

FREE E-MAIL SERVICE

Address

Password

go

Get TIME Daily delivered to your desktop every day with Microsoft Internet Explorer

June 30, 1998

U.S. Plane Fires a Missile On Iraq

An Iraqi radar station targets an Allied plane, and a U.S. F-16 responds quickly -- with deadly force. Is another showdown with Saddam on the way?
Full Story

Responding with Force: A U.S. Air Force F-16 flies over Kuwait. U.S. AIR FORCE/AP

Starr Plays the Tripp Card
The former confidante's grand jury appearance puts the squeeze on Ms. Lewinsky.

Down to Business in Shanghai
President Clinton spends some time in the city he wants the rest of China to turn into.

·Poll: Does the U.S. have the right to impose its idea of human rights on China?

·Postcards From the Middle Kingdom: TIME's Jay Branegan says President Clinton is in full campaign mode in China. But the big question is, why isn't he pressing the flesh?

Boris Duels With the Duma
If Russian president Yeltsin wants to make other Russian pols look bad, he should stop making a fool of himself first.

TV-GUIDES — decision making

2:30am **VC2 – 76**

The Jackal

Movie: Bruce Willis excels as "The Jackal," a cunning assassin who uses many disguises in this 1997 thriller. Richard Gere and Sidney Poitier costar as players from different sides of the law who unite to stop him.

3:00am **KCOP – 13**

The Untouchables

Movie: Eliot Ness (Kevin Costner) and ‘‘The Untouchables’’ take on Robert De Niro’s flamboyant Al Capone in the pulse–pounding 1987 adaptation of the popular TV series.\Sean Connery won an Oscar as the Irish beat cop who shows Ness ‘‘the Chicago way.’’\ Brian De Palma directed the feature;\David Mamet wrote the script.\And yes, film majors, the scene at Union Station was lifted directly from the

3:05am **STARZ – 25**

Grosse Pointe Blank

Movie: A razor–sharp script and a fine turn by John Cusack as a troubled hit man mark 1997’s "Grosse Pointe Blank," a dark comedy in which the assassin encounters his old flame (Minnie Driver of "Good Will Hunting") at a high–school reunion. Cusack’s sister Joan ("In and Out") is hilarious as the killer’s devoted assistant, and Alan Arkin makes the most of his small role as Cusack’s terrified the

Abstracts of papers — time saving

An Incremental Interpreter for High-Level Programs with Sensing

Giuseppe De Giacomo
Dipartimento di Informatica e Sistemistica
Universita di Roma "La Sapienza"
Via Salaria 113, 00198 Rome, Italy
degiacomo@dis.uniroma1.it

Hector Levesque
Department of Computer Science
University of Toronto
Toronto, Canada M5S 3H5
hector@cs.toronto.edu

Abstract

Like classical planning, the execution of high-level agent programs requires a reasoner to look all the way to a final goal state before even a single action can be taken in the world. This deferral is a serious problem in practice for large programs. Furthermore, the problem is compounded in the presence of sensing actions which provide necessary information, but only after they are executed in the world. To deal with this, we propose (characterize formally in the situation calculus, and implement in Prolog) a new incremental way of interpreting such high-level programs and a new high-level language construct, which together, and without loss of generality, allow much more control to be exercised over when actions can be executed. We argue that such a scheme is the only practical way to deal with large agent programs containing both nondeterminism and sensing.

Introduction

In [4] it was argued that when it comes to providing high level control to autonomous agents or robots, the notion of *high-level program execution* offers an alternative to classical planning that may be more practical in many applications. Briefly, instead of looking for a sequence of actions $\vec{a}$ such that

$$Axioms \models Legal(do(\vec{a}, S_0)) \wedge \phi(do(\vec{a}, S_0))$$

where ϕ is the goal being planned for, we look for a sequence $\vec{a}$ such that

$$Axioms \models Do(\delta, S_0, do(\vec{a}, S_0))$$

to find a sequence with the right properties. This can involve considerable search when δ is very nondeterministic, but much less search when δ is more deterministic. The feasibility of this approach for AI purposes clearly depends on the expressive power of the programming language in question. In [4], a language called ConGolog is presented, which in addition to nondeterminism, contains facilities for sequence, iteration, conditionals, concurrency, and prioritized interrupts. In this paper, we extend the expressive power of this language by providing much finer control over the nondeterminism, and by making provisions for sensing actions. To do so in a way that will be practical even for very large programs requires introducing a different style of on-line program execution.

In the rest of this section, we discuss on-line and off-line execution informally, and show why sensing actions and nondeterminism together can be problematic. In the following section, we formally characterize program execution in the language of the situation calculus. Next, we describe an incremental interpreter in Prolog that is correct with respect to this specification. The final section contains discussion and conclusions.

Off-line and On-line execution

To be compatible with planning, the ConGolog interpreter presented in [4] executes in an *off-line* manner, in the sense that it must find a sequence of actions constituting an entire legal execution of a program *before* actually executing any of them in the world.[1] Consider, for example, the following program:

Graphical maps — orienting

Textual Directions — planning

Door to Door Directions:

From:	6420 Green Valley Circle Culver City, CA
To:	4676 Admiralty Way Marina del Rey, CA

Direction	Distance
1: Start out going South on GREEN VALLEY CIR towards W CENTINELA AVE.	0.2 miles
2: Turn RIGHT onto S CENTINELA AVE.	0.5 miles
3: Turn RIGHT onto SEPULVEDA BLVD.	0.6 miles
4: Turn RIGHT onto W SLAUSON AVE.	0.3 miles
5: Take the CA-90 WEST ramp.	0.1 miles
6: Merge onto CA-90 W.	2.9 miles
7: Turn LEFT onto MINDANAO WAY.	0.3 miles
8: Turn RIGHT onto ADMIRALTY WAY.	0.0 miles

Total Distance:	4.9
Estimated Time:	11 minutes

Cliff notes — Laziness support

Cliff Notes for the Grapes of Wrath

Posted by Derek on December 02, 1997 at 11:35:43:

In Reply to: Re: I need cliff notes or a summary to TO KILL A MOCKING>> posted by kandice on September 28, 1997 at 20:40:48:

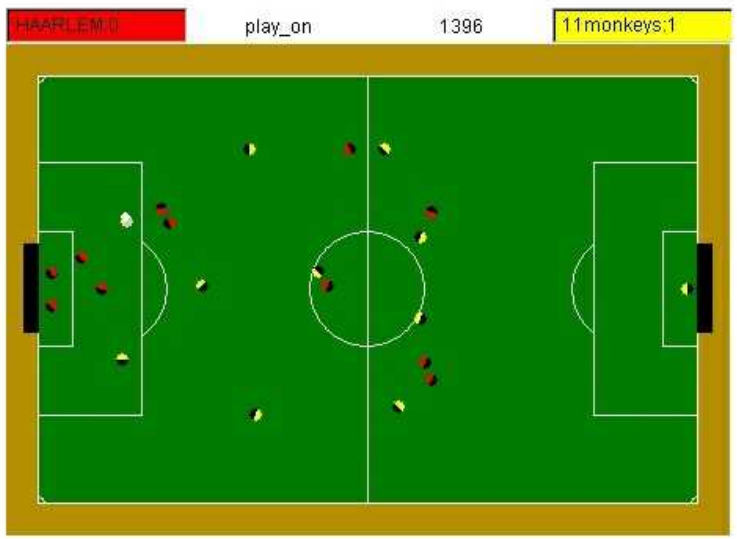
Say can you send me some cliff notes for the grapes of wrath by Wednesday December 3, 1997. I would appricate it very much and I would recomend this page to all my friends so we could ace our english tests on the grapes of wrath. PlEASE SEND ME A COPY OF THE GRAPES OF WRATH CLIFF NOTES I NEED THEM BAD!!!!!!!!!!!!!!!

Soccer Game Summaries

HAARLEM offense collapses in stunning defeat at the hands of 11monkeys!

11monkeys displayed their offensive and defensive prowess, shutting out their opponents 7-0. 11monkeys pressed the attack very hard against the HAARLEM defense, keeping the ball in their half of the field for 84% of the game and allowing ample scoring opportunities. HAARLEM pulled their defenders back to stop the onslaught, but to no avail. To that effect, 11monkeys was able to get past HAARLEM's last defender, creating 2 situations where only the goalie was left to defend the net. 11monkeys also handled the ball better, keeping control of the ball for 86% of the game. HAARLEM had a tendency to keep the ball towards the center of the field as well, which may have helped lead them to ruin given the ferocity of the 11monkeys attack.

11monkeys scored using their dribbling technique for 7 of their goals. HAARLEM did not keep a good amount of distance between their players. 11monkeys displayed some of their ball control skills. HAARLEM had their last defender bypassed 2 times for 1 goals.



- AI Agent plans summary
- Winner of prize at IJCAI (Tambe et al., 1999)

Questions

- **What kinds** of summaries do people want?
  - What are *summarizing, abstracting, gisting*,...?
- **How sophisticated** must summ. systems be?
  - Are statistical techniques sufficient?
  - Or do we need symbolic techniques and deep understanding as well?
- **What milestones** would mark quantum leaps in summarization theory and practice?
  - How do we measure summarization quality?

文档摘要定义

- 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)

文档摘要的种类

报道性文摘 informative abstracts

- 概括叙述原文献中的重要事实情报，包括研究对象、工作目的、主要结果，以及与研究性质、方法、条件、手段等有关的各种资料，在一定程度上可代替原文献。

指示性文摘 indicative abstracts

- 指明原文献的主题与内容梗概，为读者查检和选择文献提供线索。

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。

文档摘要的种类

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

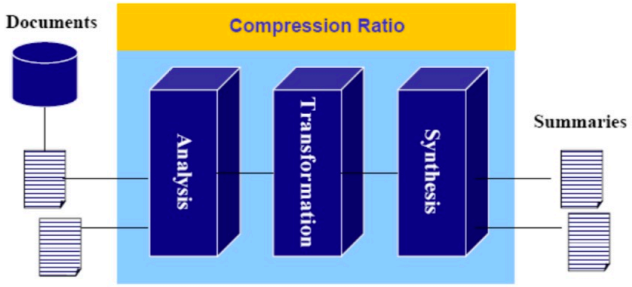
- 由文献作者以外的人员编写的文摘。

任务分类

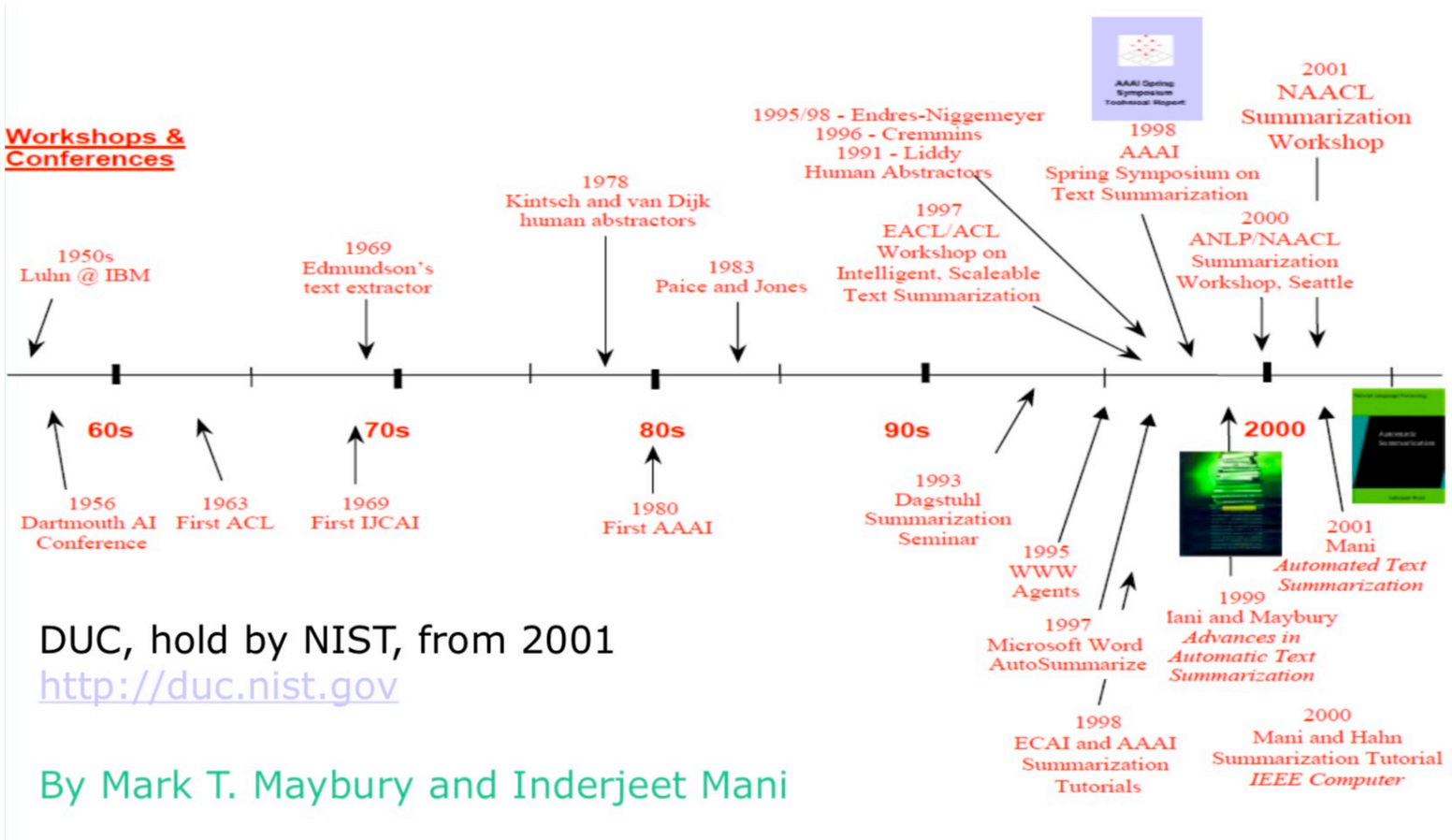
- 用户需求：
  - Generic summarization
  - User-query summarization
  - Task-oriented summarization
  - Query-based summarization
  - Etc.
- 文本性质
  - Single-document summarization
  - Multi-document summarization
- 方法
  - Extractive summarization
  - Abstractive summarization
- 训练依赖条件
  - Supervised summarization
  - Unsupervised summarization

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。



文档摘要历史



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

By Mark T. Maybury and Inderjeet Mani

文本摘要的重要性

文本自动摘要的重要性

- 我们进入一个信息爆炸的时代
  - 据IDC统计，互联网数据量已跃至ZB级别（1ZB=2^{40}GB），预计2020年达到35ZB

- 搜索引擎不能有效解决信息过载的问题
  - 相关信息过多：冗余、片面、杂质

- 移动设备的普及使用
  - 屏幕小、网络带宽低等特点需要新的信息浏览与阅读方式

Baidu新闻 新闻 网页 贴吧 知道 音乐 图片 视频 地图 百科 文库

百度新闻无线版 | 百度首页 | 登录 注册

郭美美事件 百度一下 帮助 | 高级搜索 | 设置

新闻全文 新闻标题

把百度设为首页

郭美美 百度新闻 - 人物

用户关注度(周): -50% 媒体关注度(周): -63% 媒体收录量:624687

事件聚焦 - 图片新闻 - 近期趋势 - 历史热点新闻

今日话题:你怎么看郭美美事件持续发酵? 金鹰网 2013-07-17 10:52:00

但这一事件也启示我们,就是应把一切真实的情况告诉公众。 红会称与郭美美划等号十分冤屈 你怎么看待郭美美事件持续发酵? 声明:新浪网独家稿件,转载请注明出处。 ... 3条相同新闻 - 百度快照

"把郭美美和红会画等号不公平" 北京青年报 2013-07-17 06:55:08

本报讯 中国红十字会党组书记、常务副会长赵白鸽昨日做客中新网时表示,"郭美美事件"伤害不仅仅是红会,对于目前仍有人把郭美美变成LOGO与红会画等号的行为,不... 15条相同新闻 - 百度快照

"郭美美事件应该画个句号" 南方都市报数字报 2013-07-18 03:05:30

A:一个小女孩把中国慈善界搞得天翻地覆的,这社会事实上是不正常的,我觉得郭美美事件应该个句号,希望这个事件能够尽快平息下来。另外呢,通过"郭美美事件",也... 2条相同新闻 - 百度快

云南探讨公益慈善组织改革 鼓励政府从募捐市场退出 中国新闻网 2013-07-19 01:12:23

然而,因为慈善捐款而引发的种种纠纷以及"郭美美事件"后,社会掀起一阵慈善问责风暴,中国慈事业遭遇了前所未有的诚信危机。何道峰在指出中国公益慈善组织面临巨大... 3条相同新闻 - 百度快照

#美国队长2# (13619) 全部讨论：7545560

漫威影业荣誉出品《美国队长2》，美国队长史蒂夫-罗杰斯将与黑寡妇强强联手，在华盛顿共同迎战强大、邪恶又神秘的黑暗劲敌——冬兵，他似乎来自队长的过去？！克里斯-埃文斯、斯嘉丽-约翰逊等主演，《美国队长2》4月4日3D、IMAX 3D、中国巨幕与北美同步上映！

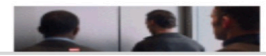
#美国队长2# 发布

主持人推荐

漫威影业 V：【《#美国队长2#》中国票房破5亿】《美国队长2》好评热映，上周末更击败《里约大冒险2》等新片，强势蝉联北美及中国内地票房冠军，中国内地总收入更势如破竹突破5亿人民币，加冕2014年引进片票房冠军！影片口碑亦持续走高，在各大影迷网站都是近期最受好评的影片！http://t.cn/8siuMag

今天 11:26 来自微博 weibo.com (118) | 转发(136) | 评论(64)

漫威影业 V：#美国队长2小花絮#电梯格斗是拍摄的第一场格斗戏。十个人与美国队长在拥挤的电梯里大战，最大的挑战就是如何将更多的动作展现于有限的空间。最终团队意识到一开始要让队长处于守势，只能在近距离使用他的拳脚，直到后来空间稍微变大，他才得以展示格斗技巧。(GIF酷炫动图)《#美国队长2#》全国热映中！



研究现状：从传统文摘到互联网摘要

- **传统文档摘要**
  - **面向新闻文档**
    - **单文档 & 多文档**
- **互联网文本摘要**
  - **面向互联网异质文本**
    - **新闻、社交媒体、学术文献、等**
- **摘要的范畴变大**

文档摘要技术回顾

- **早期论文**
  - **Luhn. The Automatic Creation of Literature Abstracts (1958)**
- **研究50多年，取得一定进展，但仍不能令人满意**
- **困难在哪里？**
  - **与人工智能、自然语言理解其他任务类似，甚至更加困难**
  - **机器写摘要 vs. 专家写摘要**
- **代表性系统**
  - **NewsInEssence by University of Michigan**
  - **NewsBlaster by Columbia University**

Columbia Newsblaster

Summarizing all the news on the Web

Tuesday, October 26, 2010
Articles from 10/23/2010 to 10/26/2010
Last update: 4:33 AM EST

Search for:

Offline summarization

Go

U.S.
World
Finance
Sci/Tech
Entertainment
Sports

View Today's Images

View Archive

About Newsblaster

About today's run

In mortgage crisis, uneasy politics (U.S., 8 articles)

" There's no question that this foreclosure issue is, in certain parts of the country, a huge issue, and anything that can be done to stop foreclosures is important said Mark Mellman , a Democratic pollster working for Reid. A wave of lawsuits from troubled borrowers who lost their homes to foreclosure could hurt the already-ailing housing market, a top banking regulator said Monday. Bair said her agency is working closely with other federal regulators to determine whether mortgage companies used flawed documents to seize homes, during remarks made at a housing finance conference sponsored by the FDIC and the Federal Reserve. Raising pressure on banks, the Federal Reserve is wading into the investigation of whether mortgage lenders cut corners and used flawed documents to foreclose on homes. Federal Reserve Chairman Ben Bernanke added weight to those efforts Monday by saying the central bank would look "intensively" at policies and procedures that might have allowed banks to seize homes improperly. SLOPPY, greedy mortgage lenders helped inflate the housing bubble of the Great Recession, now the Federal Reserve is investigating to see if the same avaricious instincts are being applied to home foreclosures.

Other stories about Mortgage, borrowers and lenders:

- Mortgage Insurance Reserves: A Lesson in Managing Risk (4 articles)
- Mortgage Web Site Has Admirable Goals, But Will It Save You Money? (4 articles)
- Time for the Government to Go Far Beyond Rate-Freeze Plan for Borrowers (4 articles)

文档摘要技术回顾

- **针对传统文档摘要任务的主要方法**
  - **抽取式**
    - **实现简单，保留完整句子，可读性良好**
    - **基于启发式规则**
      - **句子位置、句子TFIDF、线索词等**
    - **基于机器学习 (Summly采用)**
      - **句子分类、序列标注、句子排序等**
    - **目前的最优方法举例**
      - **整数线性规划(ILP)**
      - **次模函数最大化(Submodular Function Maximization)**
      - **行列式点过程(DPP)**

文档摘要技术回顾

- **针对传统文档摘要任务的主要方法**
  - **压缩式**
    - **同时进行句子抽取与压缩或融合**
    - **能有效提高ROUGE值，但会牺牲句子可读性**
  - **生成式**
    - **直接从意义表达生成摘要句子**
    - **难度大，更接近摘要的本质**
    - **目前效果不佳，但值得鼓励**

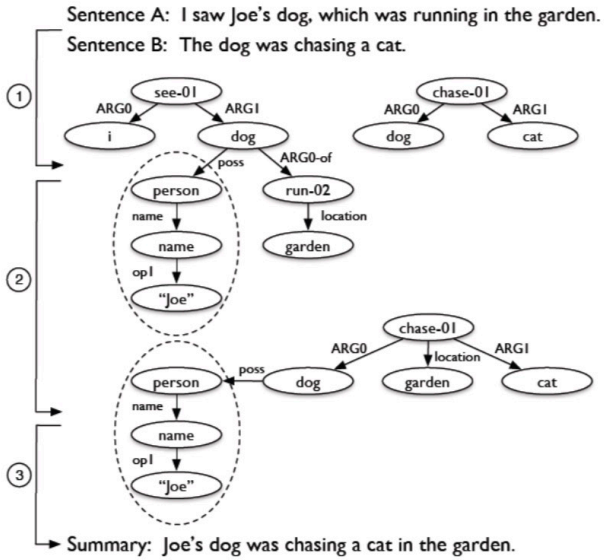


Figure 1: A toy example. Sentences are parsed into individual AMR graphs in step 1; step 2 conducts graph transformation that produces a single summary AMR graph; text is generated from the summary graph in step 3.

文档摘要技术回顾

- **单文档摘要效果理想**
  - 单文档内容紧凑，句子位置比较重要
    - Lead系统是一个很强的基准系统
  - 摘要可读性容易保证
  - 可以实用

- **多文档摘要效果不是特别理想**
  - 摘要可读性不易保证
  - 不同人对多文档中的要点有不同看法，因此对于同一摘要质量也有不同看法

文档摘要的评价方法

- 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。

Edmundson评价

- Edmundson评价
  - 属于内部评价方法
  - 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。

Edmundson评价

- 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。

评价方法

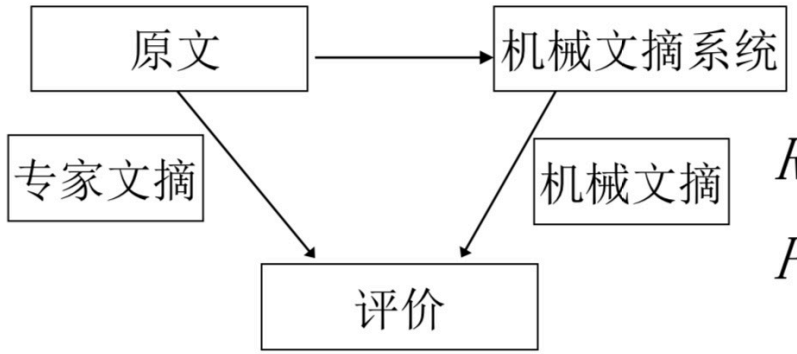
- 定义：

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）



$$Recall = N_{hm} / N_h$$

$$Precision = N_{hm} / N_m$$

ROUGE 评测

- 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

ROUGE 评测

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

- 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 评测数据

- 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.

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

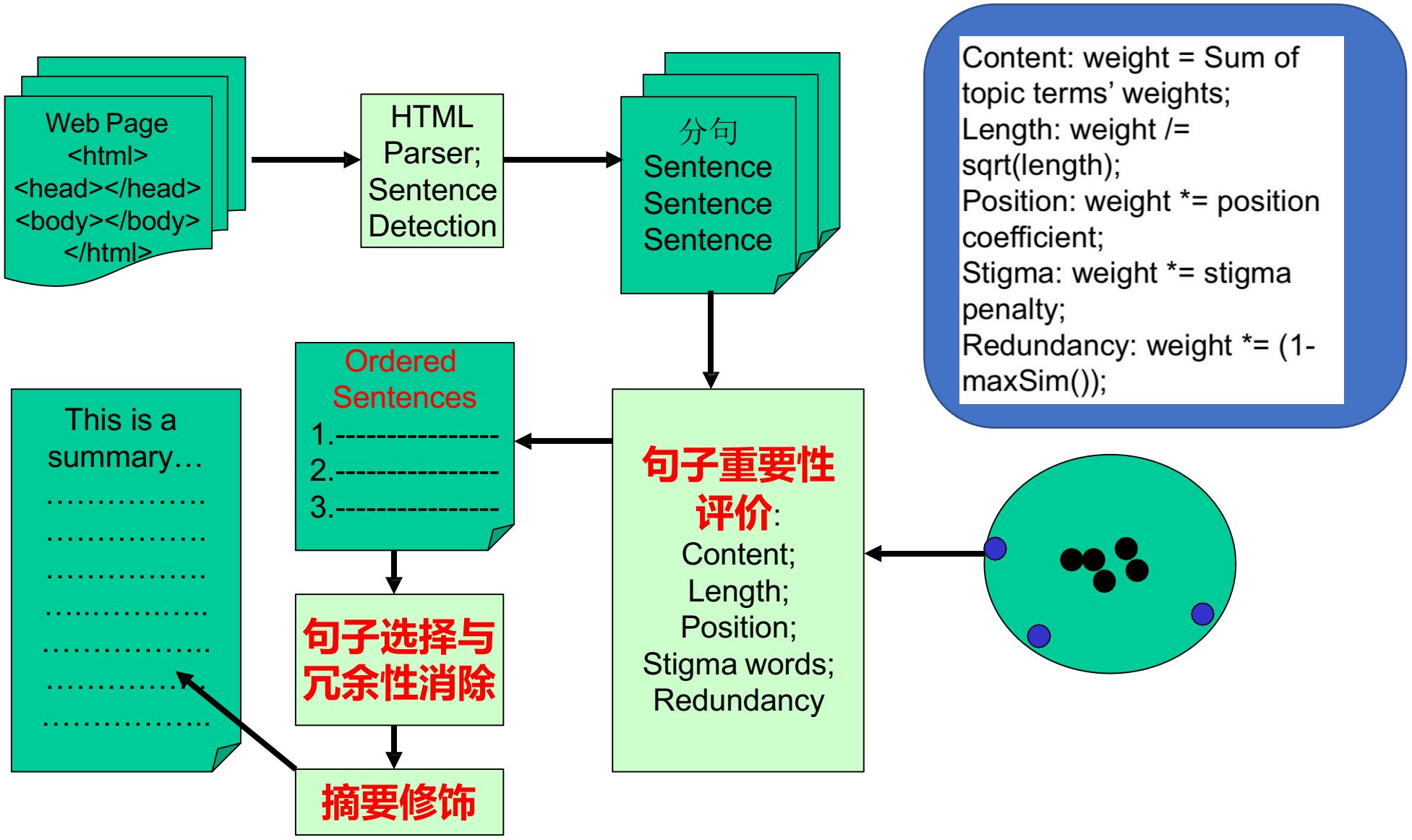
- 基于理解的自动摘要常包含语法分析、语义分析、信息提取和文摘生成，作者文摘应属于此。
- 研究表明：理解首先应着重篇章理解、段落理解，也就是理解应该是分层的，高层理解比低层理解更为重要。

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

Summarization Review



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)

Sentence Importance--Score

- Statistical features
  - Tf
  - Tf-idf
- Linguistic features
  - Location
    - 段落的位置
    - 句子在段落中的位置
  - Semantic
- Integrative features

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

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(重复) 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

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

- 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%

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%)

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
- 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。

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

MMR相关方法

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。

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。

MMI-MS

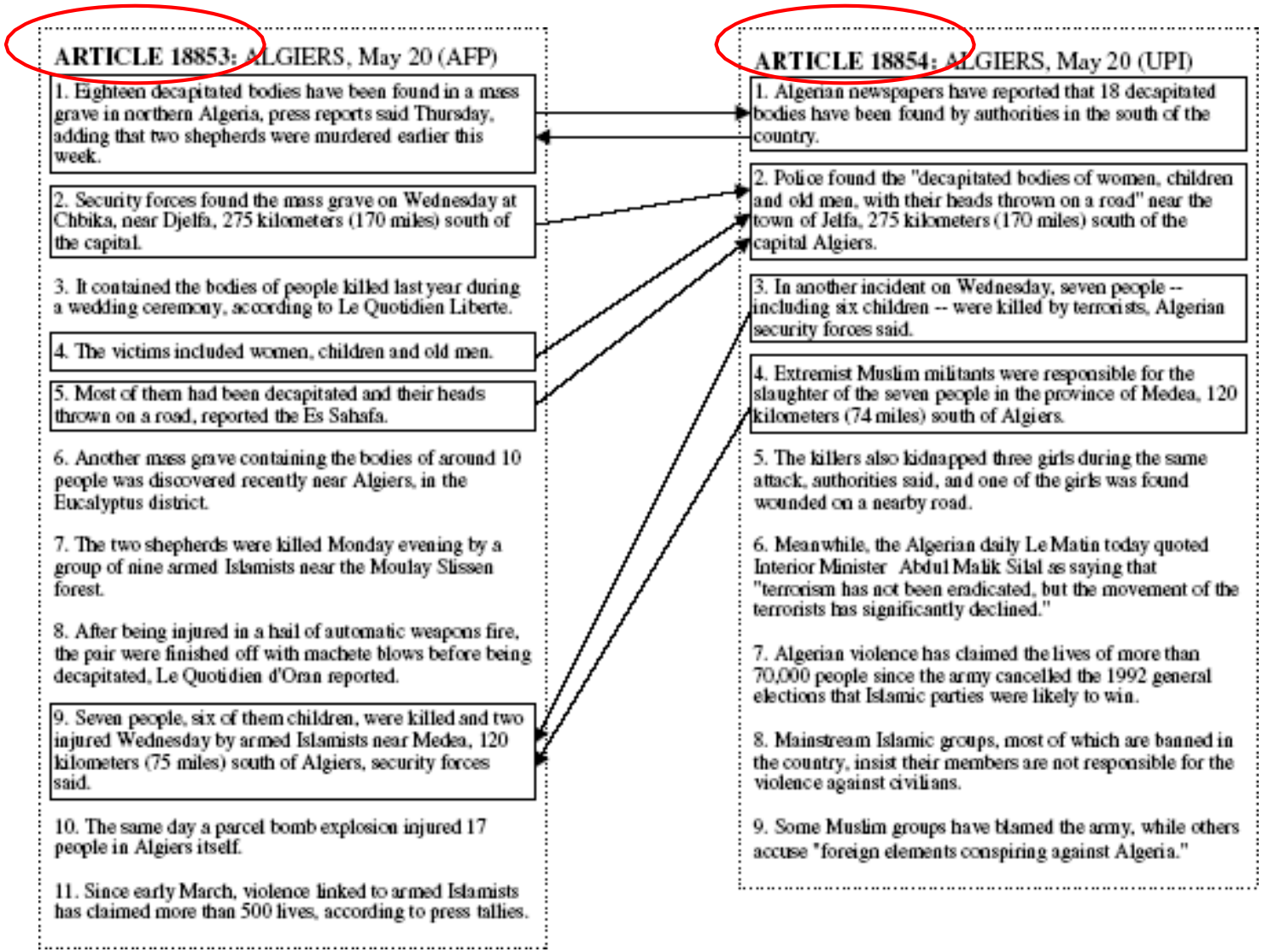
- 日本横滨国立大学开发的一个多文档自动文摘系统将MMR技术和IGR (Information Gain Ratio)技术结合起来，称为MMI-MS (Maximal Marginal Importance – Multi-Sentence)来选取文摘句。

多文档摘要

多文档摘要

- multi-document summarization
- 意义：服务于话题检测、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；
- MEAD : http://www.summarization.com/mead/
  - MEAD is a public domain portable multi-document summarization system

多文档摘要难点


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."

Clustering based Algorithm

Clustering based 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。

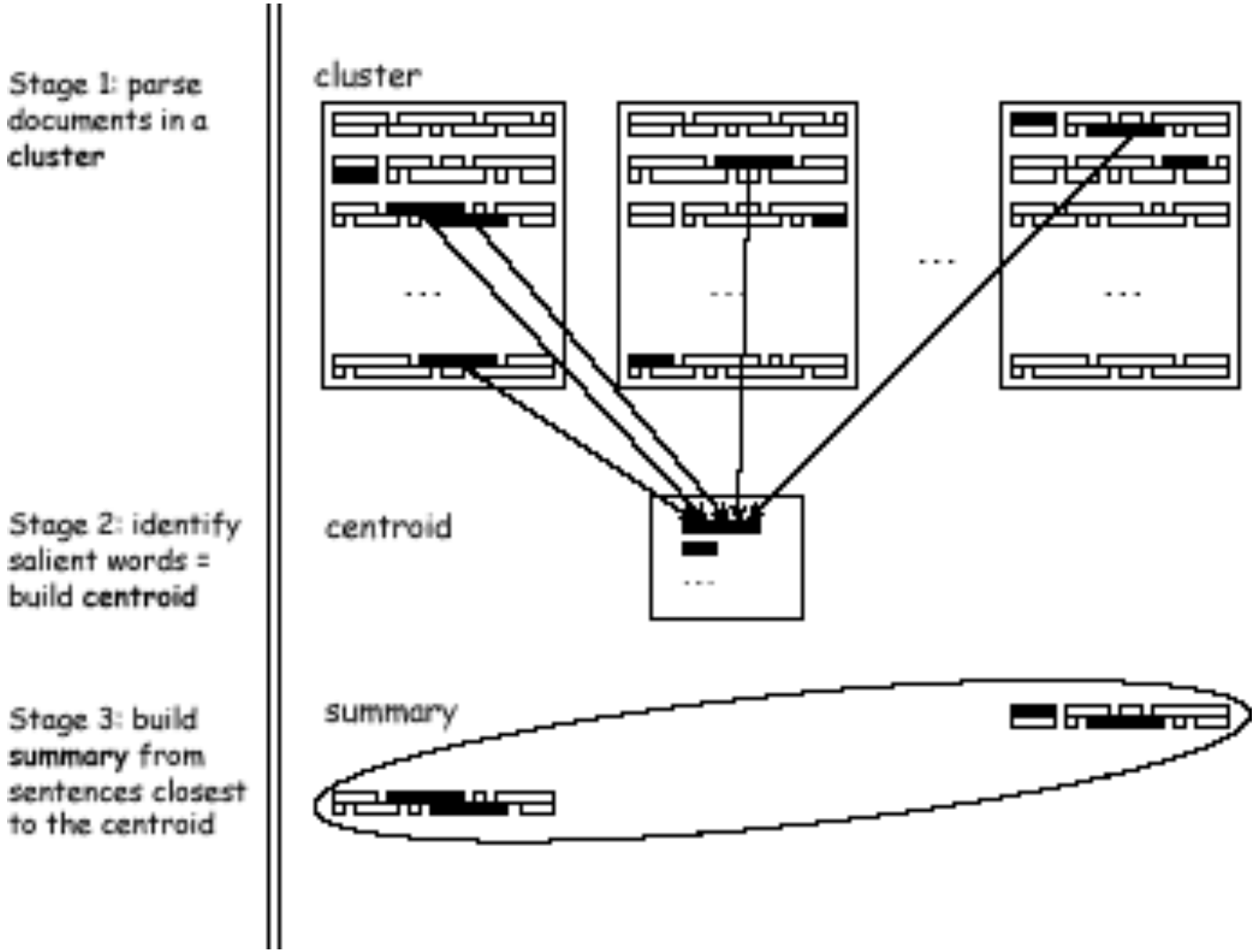
Clustering based Algorithm

- 方法
  - 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

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。
  - 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，例如：文章首句通常很重要。

Centroid-based summarization (MEAD)



CollabSum: Exploiting Multiple Document Clustering for Collaborative Single Document Summarizations

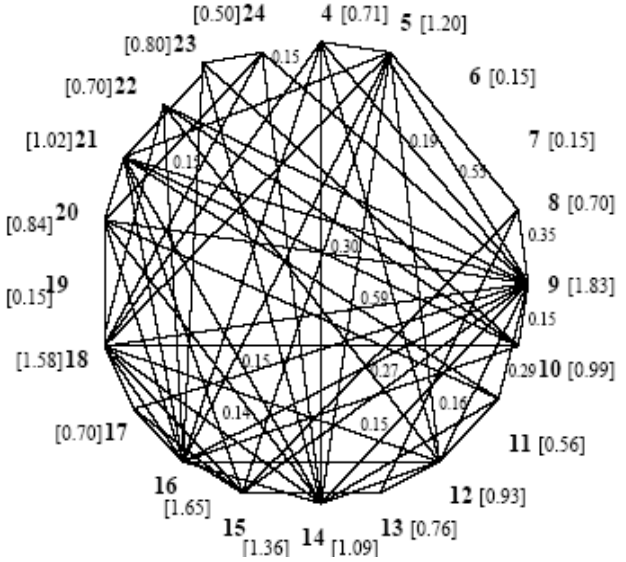
(SIGIR 2007)

Motivation

- Single document summarization
  - 只利用了单文档内部的信息；
  - 假定文档之间是相互独立的；
  - 没有考虑文档之间的相互作用；
- Motivation
  - 文档类簇中文档之间的相互影响；
  - 对句子评价过程中相关文档能够提供知识和线索；
- Collaborative summarization
  - 利用多个文档改善单文档摘要；

Related Works

- Graph-ranking based methods
  - Graph-ranking
    - TextR ank [Mihalcea and Tarau 2004, 2005]
    - LexP ageRank [ErKan and Radev, 2004]:
  - 只利用文档内句子之间的相互投票来评价句子;



Collaborative Summarization(1/7)

Underlying Idea

- Underlying idea:
  - 利用文档类簇中多个相关文档所包含的额外知识和更多线索，来评价和选择给定单文档中的句子;
- CollabSum can integrate any specific summarization method, such as graph-ranking method
  - The within-document relationships (local information)
  - The cross-document relationships (global information)

Collaborative Summarization(2/7)

Framework

- Document clustering
  - Produce the cluster context
  - Basis of collaborative summarization
  - Popular clustering algorithms
    - K-means, agglomerative, divisive, etc.
- Collaborative summarization of each document in a specified cluster context *D*
  - Global affinity graph (亲和图) building
  - Informativeness score computation;
  - Within-document redundancy removing;

Cluster level

Document level

Collaborative Summarization(3/7)

Global Affinity Graph Building

- Sentence set S
  - n sentences in document set D
- Affinity Graph G
  - sentences as nodes
  - Edge weight: cosine similarity between sentences

$$M_{i,j} = \begin{cases} sim_{\text{Cosine}}(s_i, s_j), & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}$$

$$\tilde{M}_{i,j} = \begin{cases} M_{i,j} \Big/ \sum_{j=1}^{n} M_{i,j}, & \text{if } \sum_{j=1}^{n} M_{i,j} \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

Collaborative Summarization(4/7)

Global Affinity Graph Building

- Global affinity graph G
  - Contain all sentence links;
  - $\mathrm{M} = (M_{i,j})_{n \times n}$ -> $\tilde{\mathbf{M}}$
- Within-document affinity graph G_{intra}
  - Contain only within-document sentence links;
  - $\mathrm{M}_{\text{intra}}$ -> $\tilde{\mathbf{M}}_{\text{intra}}$
- Cross-document affinity graph G_{inter}
  - Contain only cross-document sentence links;
  - $\mathrm{M}_{\text{inter}}$ -> $\tilde{\mathbf{M}}_{\text{inter}}$
- $\mathrm{M}=\mathrm{M}_{\text{intra}}+\mathrm{M}_{\text{inter}}$

Collaborative Summarization(5/7)

Informativeness Score Computation

- Use Pagerank algorithm
  - Random walk model
  - Iterative form
- Based on global affinity graph G

$$IFScore_{all}(s_i) = d \cdot \sum_{all\, j \neq i} IFScore_{all}(s_j) \cdot \tilde{M}_{j,i} + \frac{(1-d)}{n}$$

$$\vec{\lambda} = d\tilde{\mathbf{M}}^{T}\vec{\lambda} + \frac{(1-d)}{n}\vec{e} \qquad \vec{\lambda} = [IFScore_{all}(s_i)]_{n\times 1}$$

Collaborative Summarization(6/7)
Informativeness Score Computation

- Based on G_{intra}

$$IFScore_{intra}(s_i) = d \cdot \sum_{all\ j \neq i} IFScore_{intra}(s_j) \cdot (\tilde{M}_{intra})_{j,i} + \frac{(1-d)}{n}$$

- Based on G_{inter}

$$IFScore_{inter}(s_i) = d \cdot \sum_{all\ j \neq i} IFScore_{inter}(s_j) \cdot (\tilde{M}_{inter})_{j,i} + \frac{(1-d)}{n}$$

- Linear combination

$$IFScore(s_i) = \lambda \cdot IFScore_{intra}(s_i) + (1-\lambda) \cdot IFScore_{inter}(s_i)$$

Collaborative Summarization(7/7)

Within-Document Redundancy Removing

- Remove redundancy for each single document, e.g. d_k
  - m sentence in d_k ($m<n$)
  - Greedy algorithm(贪心算法), similar to MMR

$$\mathbf{M}_{d_k} = (M_{d_k})_{m \times m} \rightarrow \tilde{\mathbf{M}}_{d_k}$$

$$ORScore(s_j) = ORScore(s_j) - (\tilde{M}_{d_k})_{j,i} \cdot IFScore(s_i)$$

Experiments and Results (1/11) Experimental Setup

- Clustering algorithms：
  - Gold Clustering
  - Agglomerative (AverageLink) Clustering
  - Agglomerative (CompleteLink) Clustering
  - Divisive Clustering
  - KMeans Clustering
  - Random1, Random2, Random3 Clustering
- Data sets
  - DUC2001 task 1 + DUC2002 task 1
- Evaluation metrics:
  - ROUGE-1, ROUGE-2, ROUGE-W

Experiments and Results (2/11) Experimental Setup

- Summary of data sets：

	DUC 2001	**DUC 2002**
Task	Task 1	Task 1
Number of documents	309	567
Number of clusters	30	59
Data source	TREC-9	TREC-9
Summary length	100 words	100 words

Experiments and Results (3/11) Experimental Setup

- Summarization methods
  - IntraLink (**baseline**)
    - $IFScore(s_i) = IFScore_{intra}(s_i)$
  - InterLink
    - $IFScore(s_i) = IFScore_{inter}(s_i)$
  - UnionLink
    - $IFScore(s_i) = 0.5*IFScore_{intra}(s_i) + 0.5*IFScore_{inter}(s_i)$
  - UniformLink
    - $IFScore(s_i) = IFScore_{all}(s_i)$
- Summarization systems
  - Clustering algorithm + summarization method
  - E.g. UniformLink (KMeans)

Experiments and Results (4/11) Clustering Results

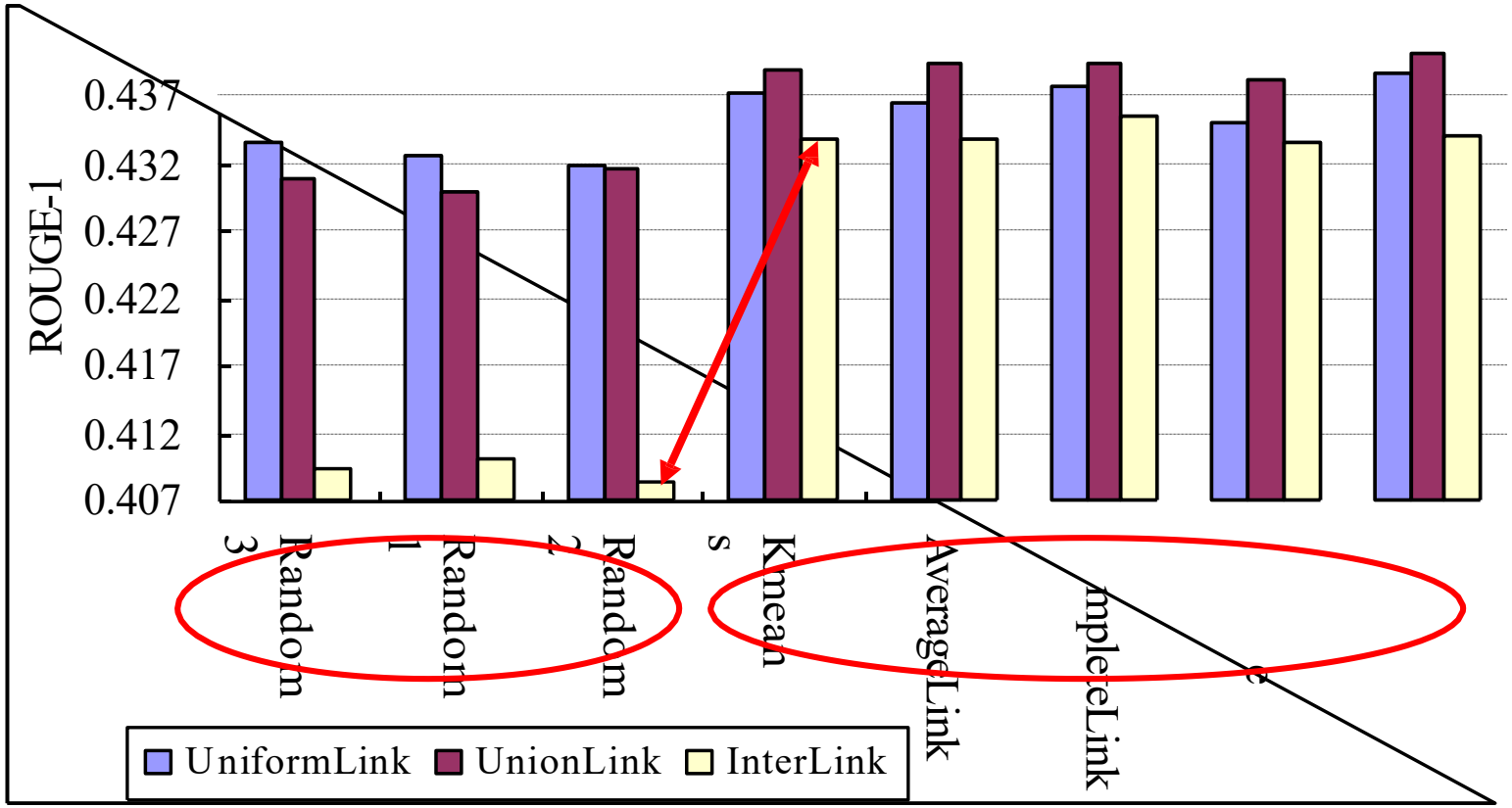
Clustering Algorithm	F-Measure	
	DUC2001	DUC2002
Gold	1.000	1.000
CompleteLink	0.907	0.799
AverageLink	0.877	0.752
Divisive	0.924	0.752
K-Means	0.866	0.722
Random1	0.187	0.168
Random2	0.189	0.168
Random3	0.183	0.167

Experiments and Results (5/11)

System	ROUGE-1	ROUGE-2	ROUGE-W
UnionLink (Gold)	0.44038	0.16229	0.13678
UnionLink (AverageLink)	0.43950	0.16108	0.13679
UnionLink (CompleteLink)	0.43947	0.16172	0.13701
UnionLink (KMeans)	0.43895	0.16054	0.13623
UniformLink (Gold)	0.43890	0.16213	0.13676
UnionLink (Divisive)	0.43832	0.15988	0.13598
UniformLink (CompleteLink)	0.43777	0.16097	0.13646
UniformLink (KMeans)	0.43726	0.15990	0.13612
UniformLink (AverageLink)	0.43651	0.15989	0.13592
InterLink (CompleteLink)	0.43556	0.15993	0.13547
UniformLink (Divisive)	0.43524	0.15846	0.13522
InterLink (Gold)	0.43422	0.15872	0.13506
IntraLink	0.43407	0.15696	0.13629

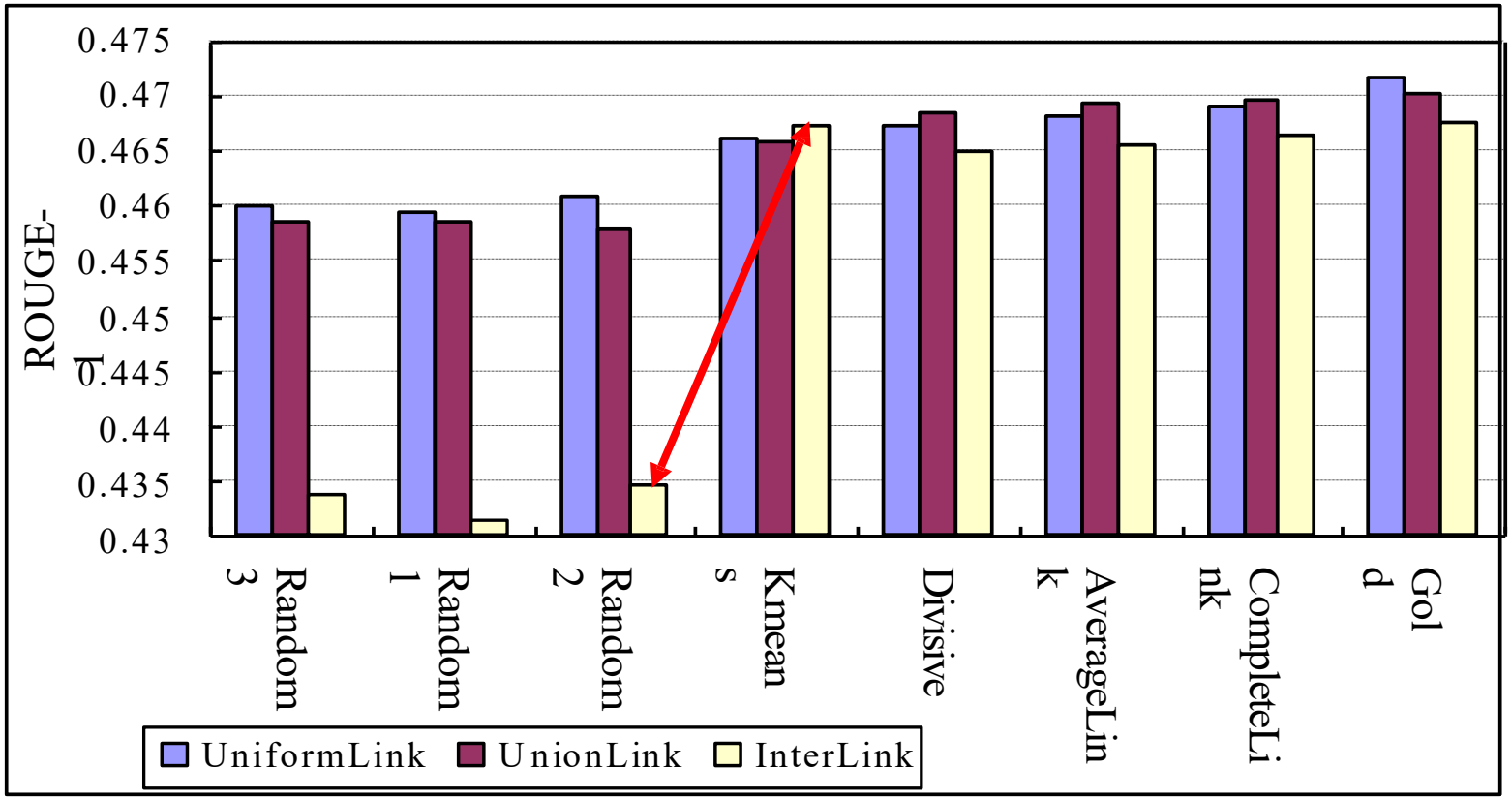
Experiments and Results (6/11)

- ROUGE-1 vs. clustering algorithm on DUC2001



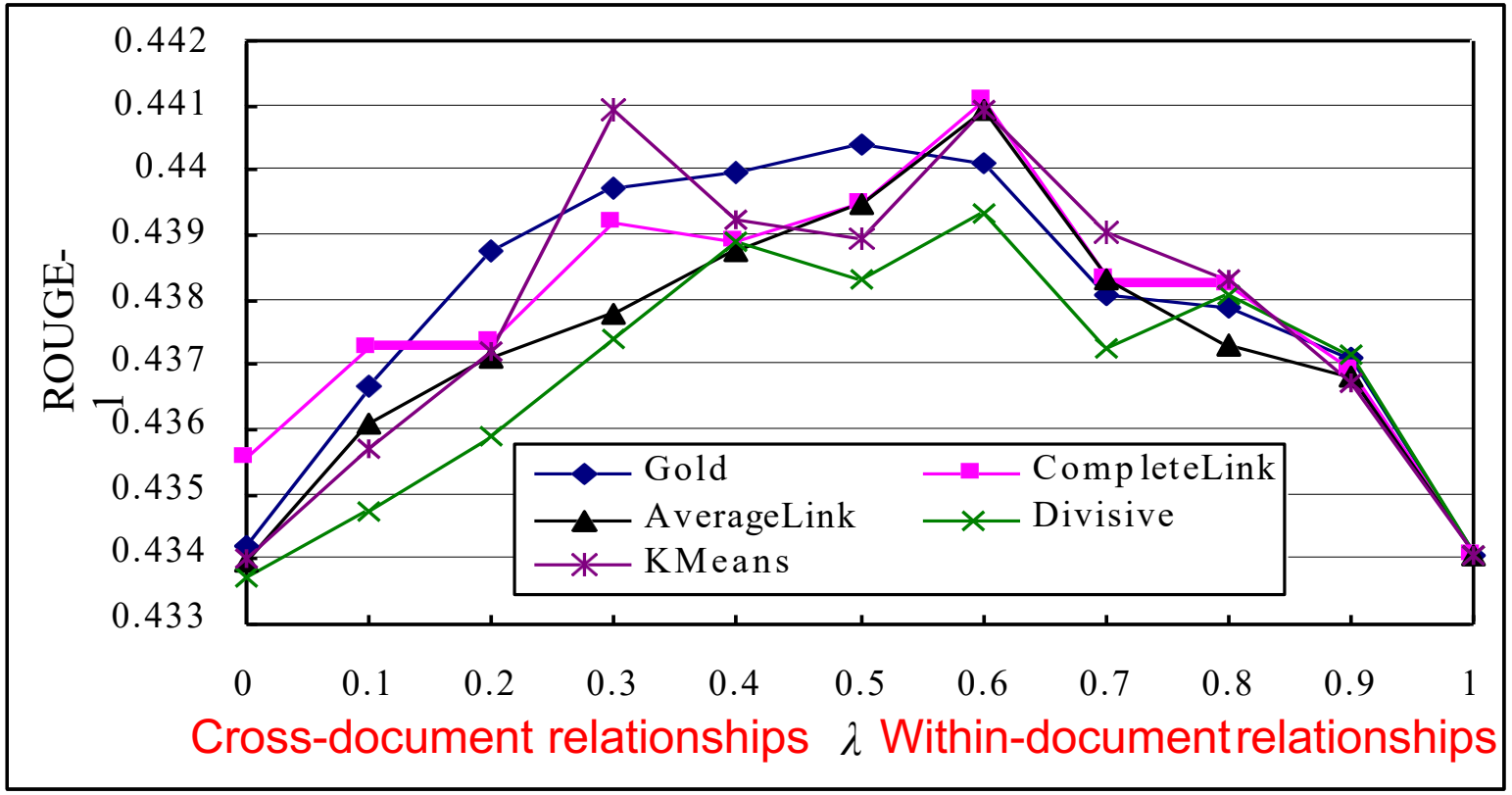
Experiments and Results (7/11)

- ROUGE-1 vs. clustering algorithm on DUC2002



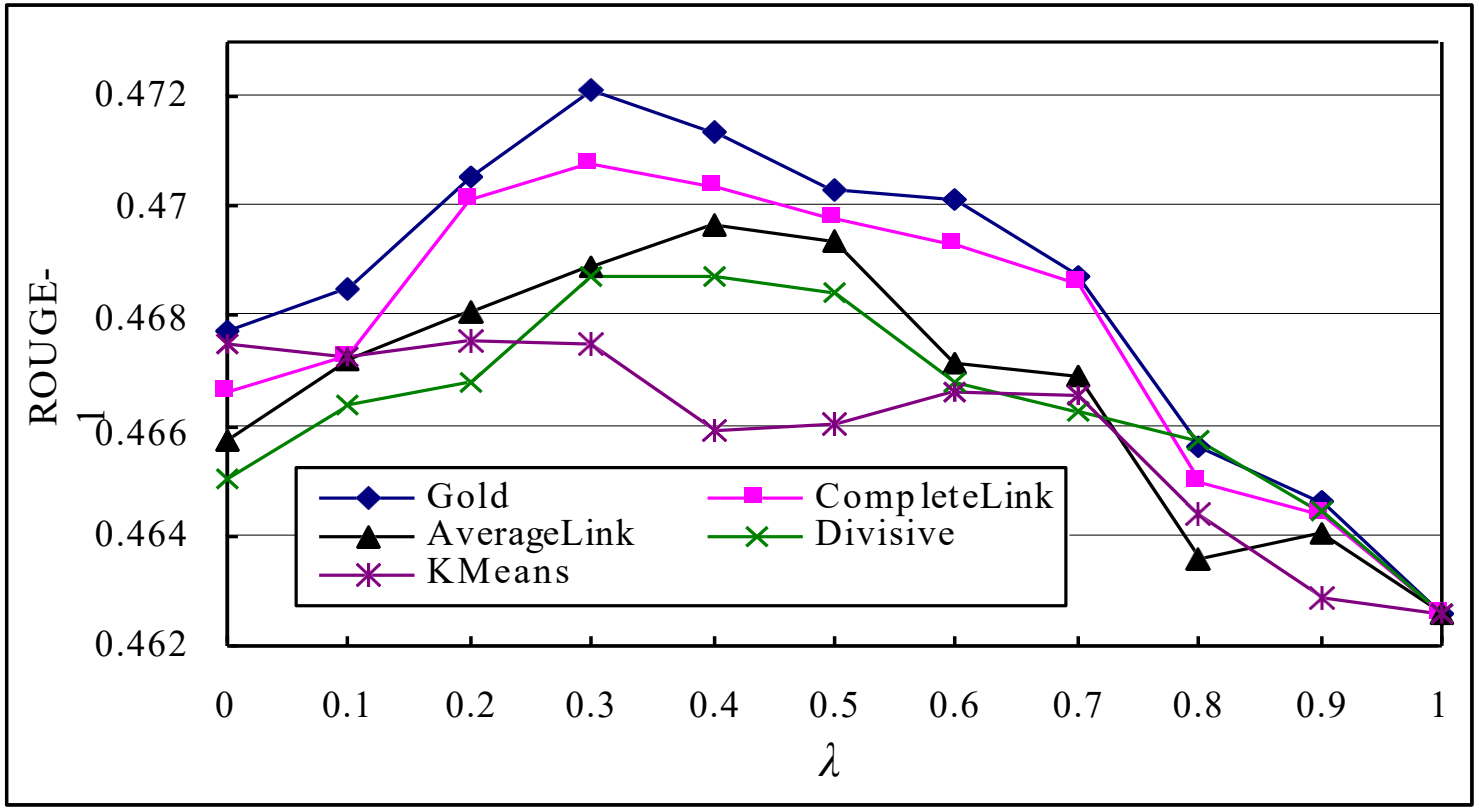
Experiments and Results (8/11)

- ROUGE-1 vs. λ on high-quality clusters of DUC2001



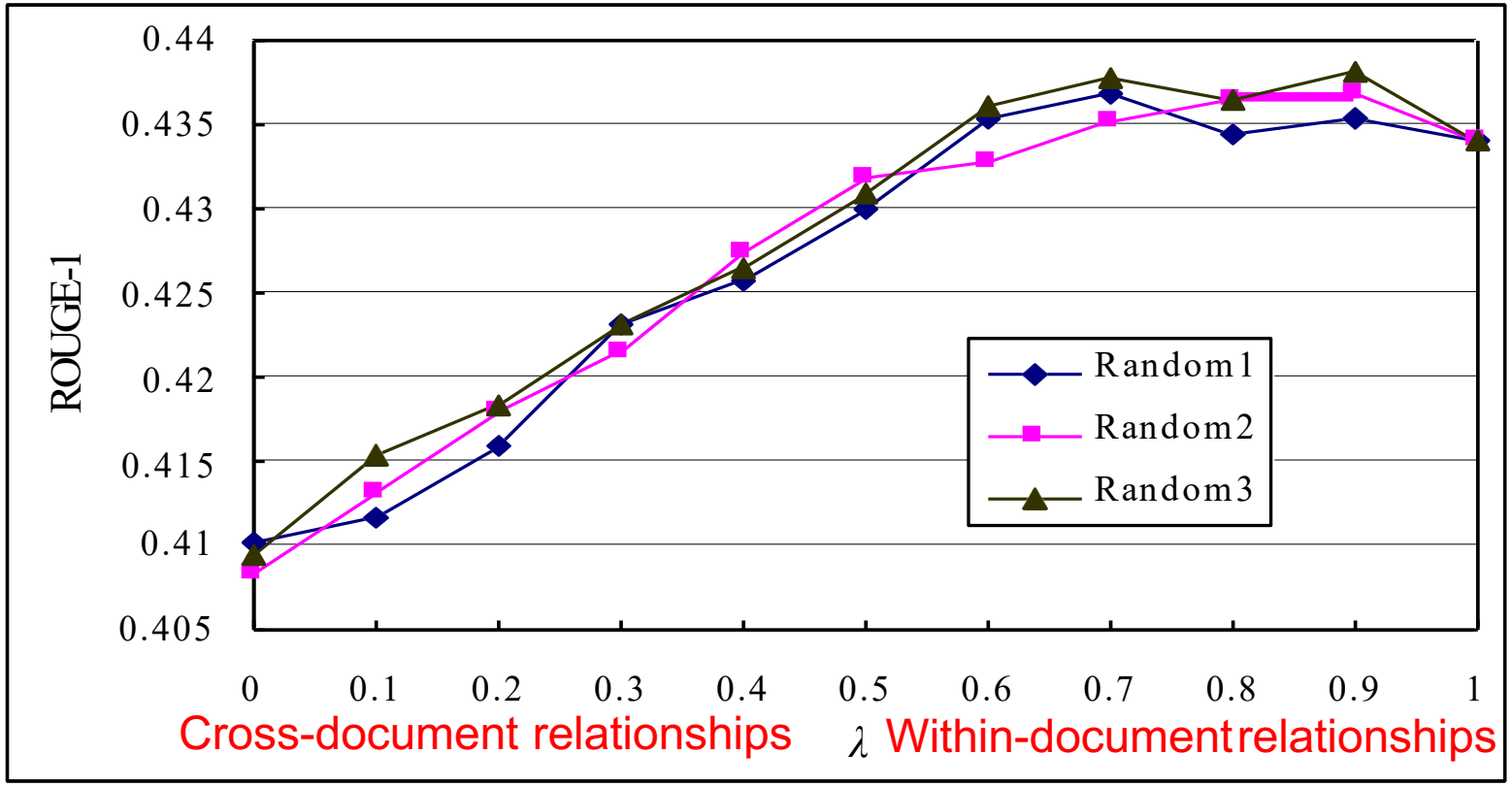
Experiments and Results (9/11)

- ROUGE-1 vs. λ on high-quality clusters of DUC2002



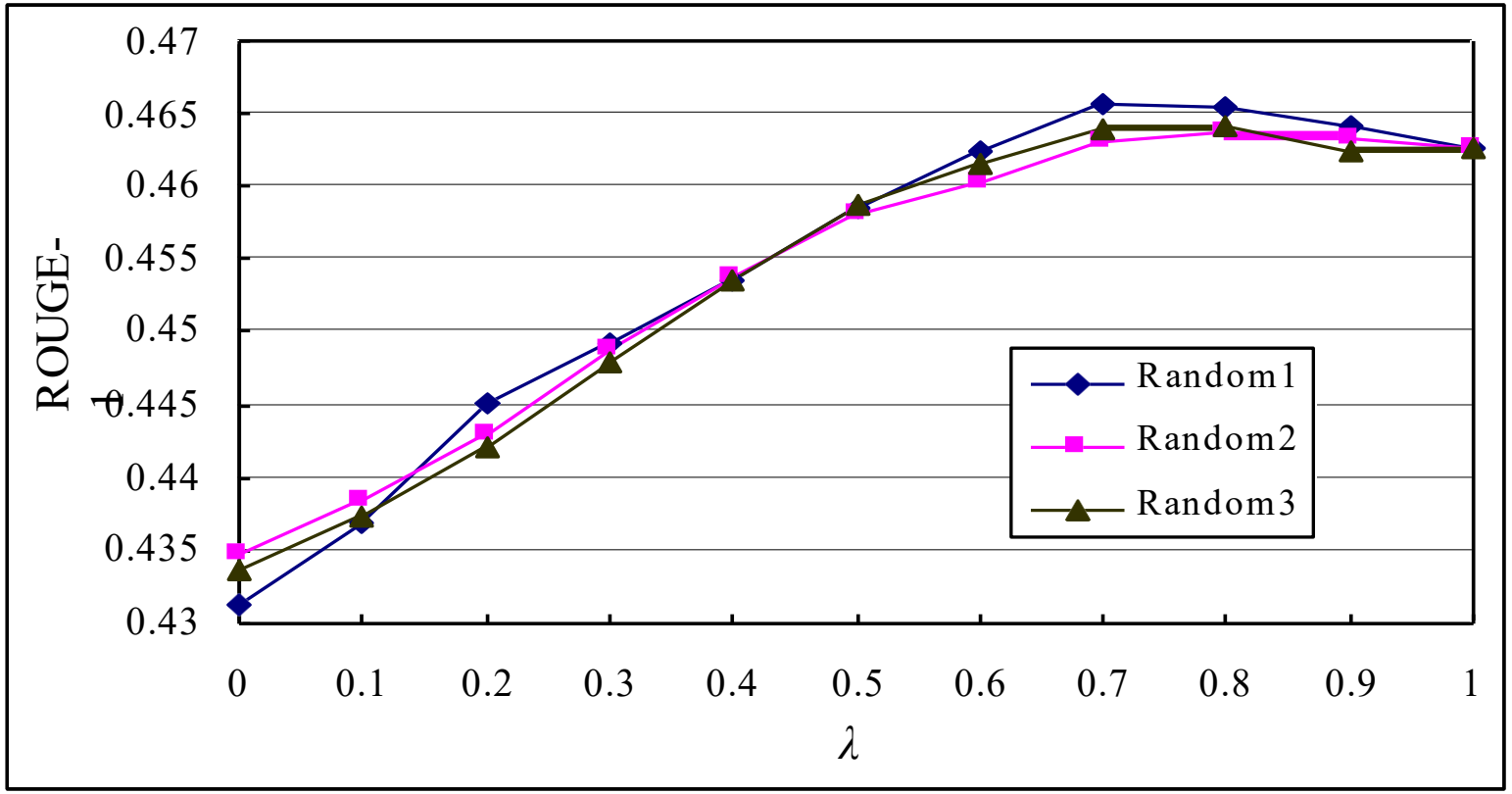
Experiments and Results (10/11)

- ROUGE-1 vs. λ on low-quality clusters of DUC2001



Experiments and Results (11/11)

- ROUGE-1 vs. λ on low-quality clusters of DUC2002



Conclusions

- Conclusions:
  - 合理的文档类簇下，跨文档之间的句子关系能够有效改善单文档摘要效果;
  - 不合理的文档类簇下，跨文档之间的句子关系会损害单文档摘要效果;
  - 现有大部分主流的聚类算法能够获得合理的文档类簇。

Multi-Document Summarization Using Cluster-Based Link Analysis

(SIGIR 2008)

Motivation

- 文档集合通常包含多个子主题，子主题具有不同的重要性
- 重要子主题类簇中的句子应该比不重要子主题类簇中的句子获得更高的评价；
- 给定一个子主题类簇，其中包含的重要句子应该比其他句子获得更高的评价；

The basic MRW model (1/3)

- MRW : Markov Random Walk Model
- Idea: PageRank of sentences
- $G=(V, E)$

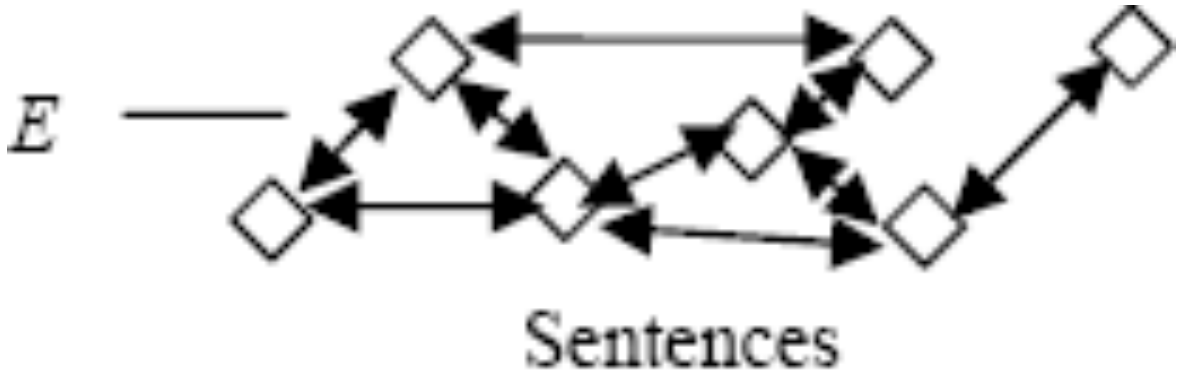


Figure 1: One-layer link graph

The basic MRW model (2/3)

$$f(i \rightarrow j) = sim_{\cos ine}(v_i, v_j) = \frac{\vec{v}_i \cdot \vec{v}_j}{|\vec{v}_i| \times |\vec{v}_j|}$$

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

$$\tilde{M}_{i,j} = p(i \rightarrow j)$$

$$SenScore(v_i) = \mu \cdot \sum_{all\, j \neq i} SenScore(v_j) \cdot \tilde{M}_{j,i} + \frac{(1-\mu)}{|V|}$$

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

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 $c_j=clus(v_i)\}$

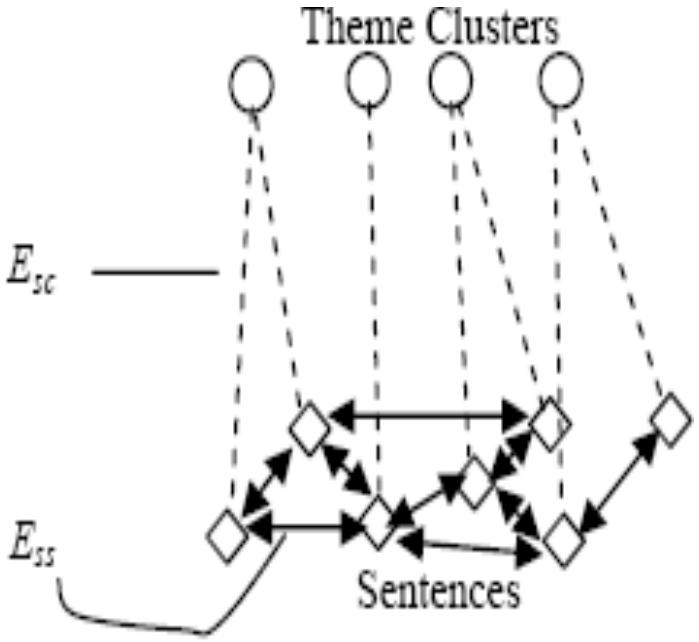


Figure 2: Two-layer link graph

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;
  - $G^{\#}=<V_s, V_c, E_{sc}>$
  - $V_s=V=\{v_i\}$ $V_c=C=\{c_j\}$
  - $E_{sc}=\{e_{ij}|v_i \in V_s, c_j \in V_c\}$

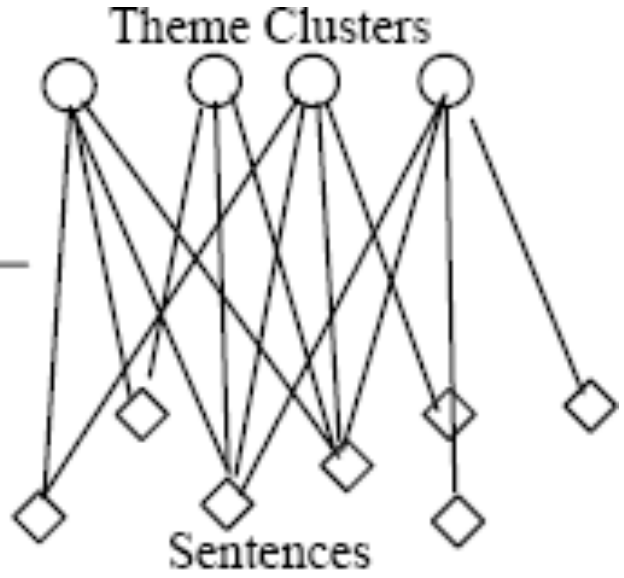


Figure 3: Bipartite link graph

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)

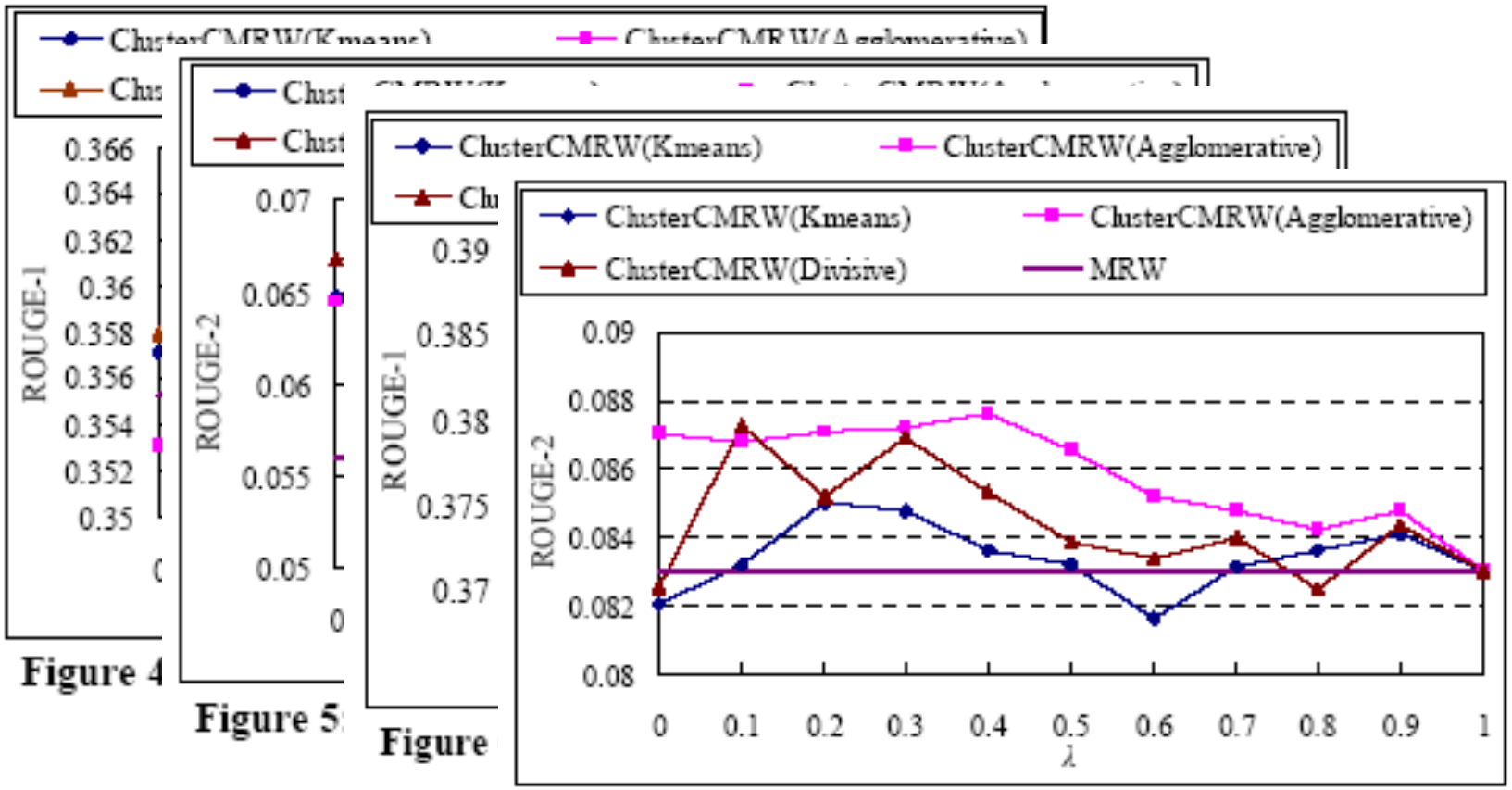


Figure 4

Figure 5

Figure

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

Experiments and Results (4/4)

- ClusterCMRW vs. ClusterHITS

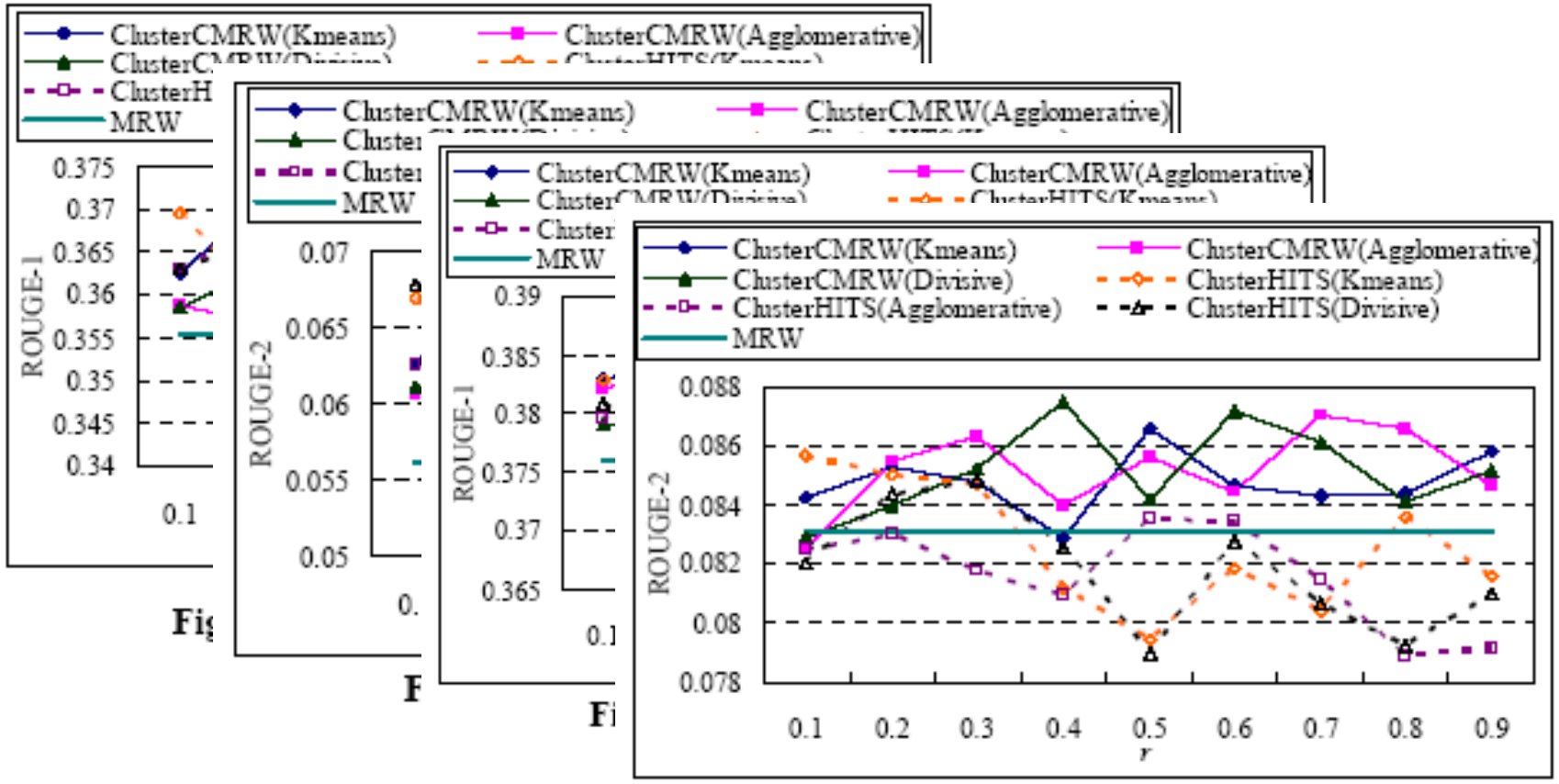
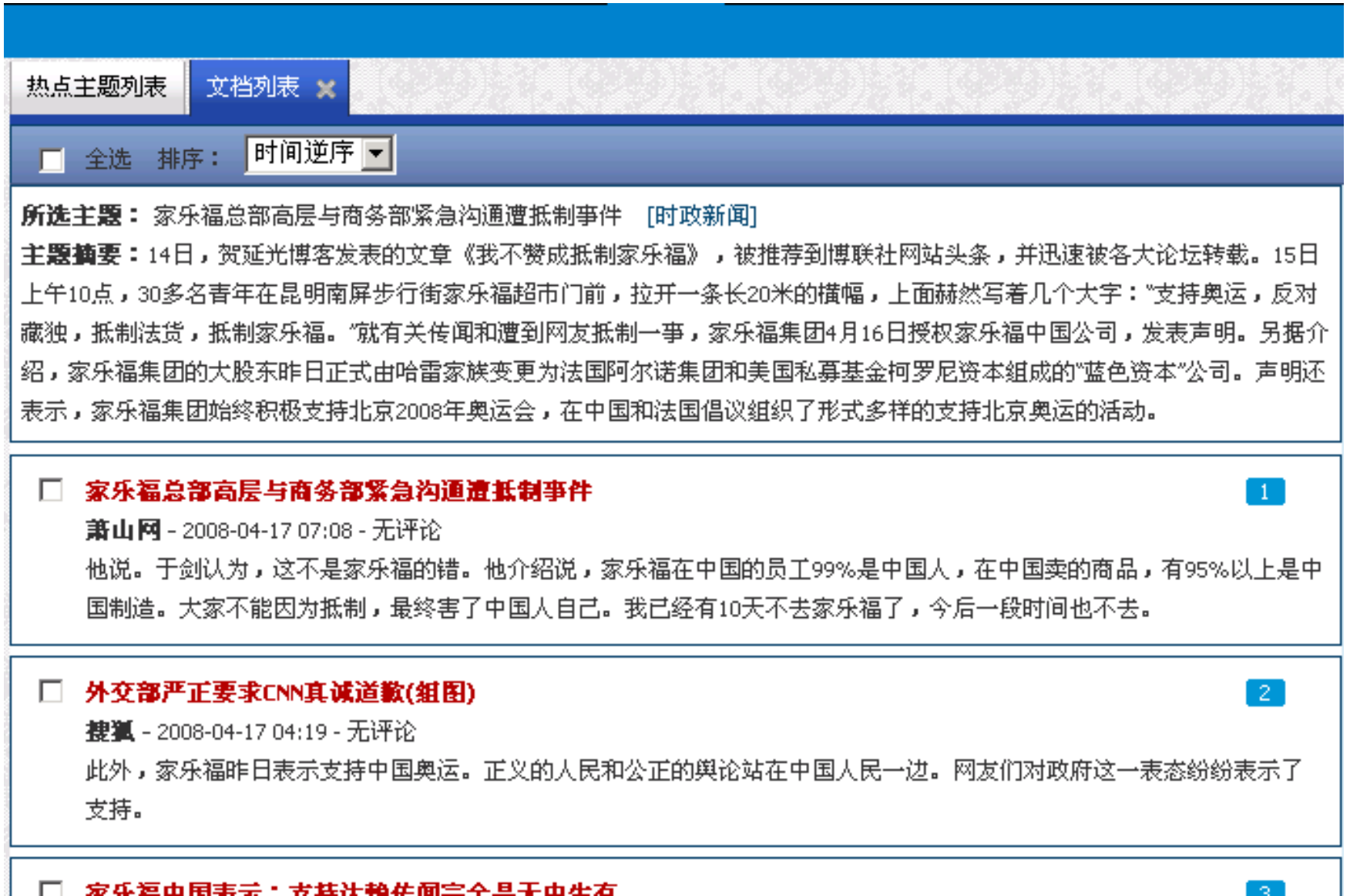


Figure 11: ROUGE-2 vs. *r* on DUC2002

Conclusions

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

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)

Next: 文档摘要进展+聚类/分类

- 任昭春
- Zhaochun.ren@sdu.edu.cn