

自我來黃州已過三寒
食年、欲惜春、意不
容惜今年又苦雨、月社
簫瑟、河海崇、花泥
污、遊支雪、閣中偷負
多、夜半、具有力、何殊、少
年、病、起、頭、白
春江欲入户、雨勢未
止、雨、小屋如漁舟、濛
水雲裏、空庭竟寒、寒
破、竈、燒、酒、華、那
知是寒食、但見烏
銜、帛、天門深
九重、黃髮在、萬里、遠、嶺
哭、淦、窮、所、不、吹、不
起

右黃州寒食二首

信息检索

Information Retrieval

教师：孙茂松

Tel: 62781286

Email: sms@tsinghua.edu.cn

TA：胡锦涛

Email: hu-jy21@mails.tsinghua.edu.cn

郑重声明

- 此课件仅供选修清华大学计算机系本科生课《信息检索》的学生个人学习使用，所以只允许学生将其下载、存贮在自己的电脑中。未经孙茂松本人同意，任何人不得以任何方式扩散之（包括放到9#服务器上）。否则，由此可能引起的一切涉及知识产权的法律责任，概由该人负责。
- 此课件仅限孙茂松本人讲课使用。除孙茂松本人外，凡授课过程中，PTT文件显示此《郑重声明》之情形，即为侵权使用。



第三章 文本分析及自动标引 (Part 3)

3.4 Thesaurus及term自动关联



汉语相关语义资源：

● 董振东教授的“知网”（HowNet）

全部资源可下载

<https://github.com/thunlp/OpenHowNet>

（注：以下几张相关Slides取自《知网》官方网站，特此致谢）



作者简介

知网简介

理论与实践

知网论坛

下载中心

相关文章

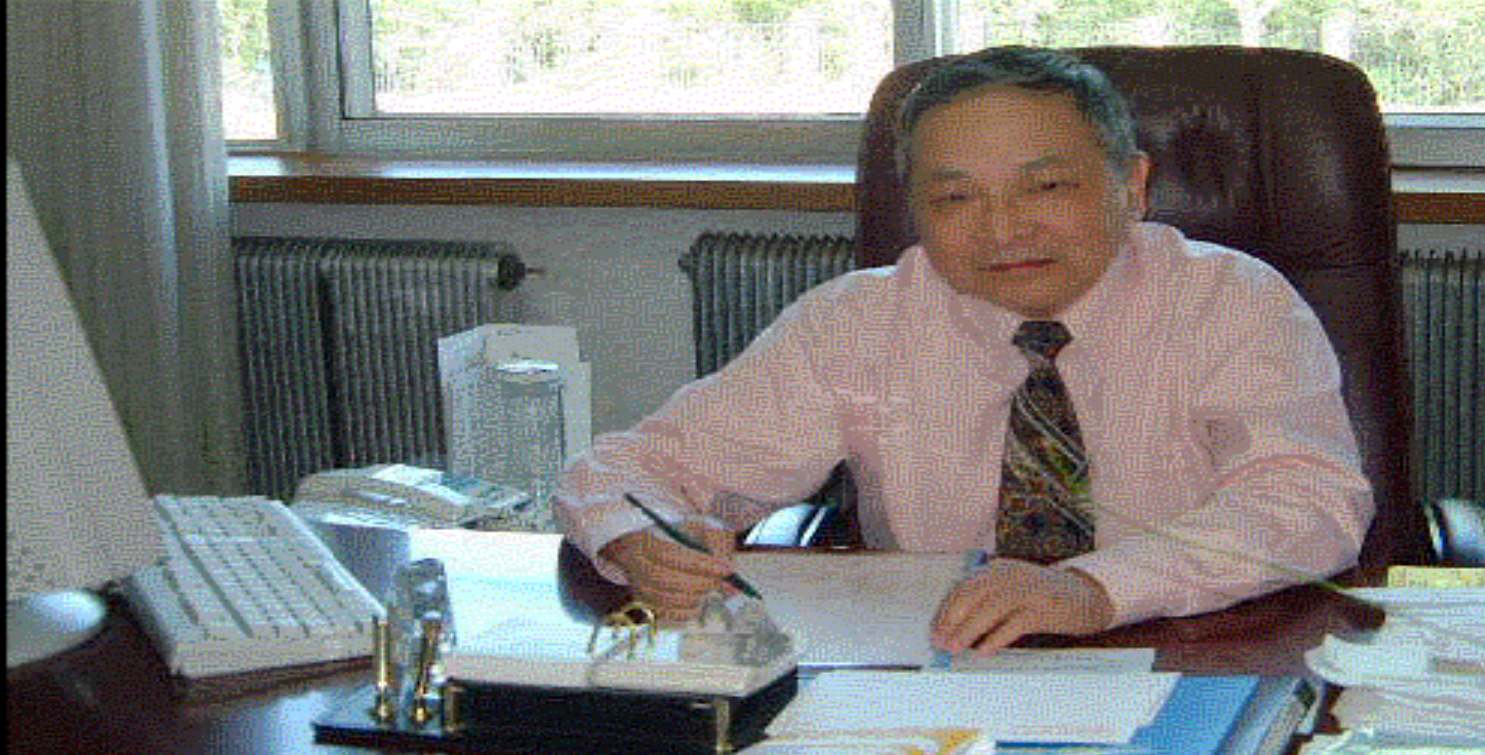
相关网站

清华论坛

102846

2006年10月23日
星期一

2006年10月23日
星期一



董振东 现任

中科院计算机语言信息中心语言知识研究室	主任
中国中文信息学会	常务理事
上海交通大学	兼职教授
北方软件学院	兼职教授

董振东 简历

1981.12 — 1989.08	军事科学院研究员，机器翻译研究组	组长
1983.12 — 1989.10	五国机器翻译国际合作项目	中方技术负责人
1989.08 — 1991.06	中国软件公司语言工程实验室	主任
1992.06 — 1993.11	日本言语研究所	主任研究员
1993.11 — 1997.01	新加坡国立大学系统科学研究院	研究员

曾担任国家“七五”机器翻译科技攻关项目
曾担任国家“八五”中文信息处理905平台工程项目

主要负责人
总体组负责人

3.4 Thesaurus及term自动关联

- (a) 上下位关系（由概念的主要特征体现，请参看《知网管理工具》）
- (b) 同义关系（可通过《同义、反义以及对义组的形成》获得）
- (c) 反义关系（可通过《同义、反义以及对义组的形成》获得）
- (d) 对义关系（可通过《同义、反义以及对义组的形成》获得）
- (e) 部件-整体关系（由在整体前标注%体现,如“心”“CPU”）
- (f) 属性-宿主关系（由在宿主前标注&体现，如“颜色”“速度”）
- (g) 材料-成品关系（由在成品前标注?体现，如“布”“面粉”）
- (h) 施事/经验者/关系主体-事件关系（由在事件前标注*体现,如“医生”“雇主”）
- (i) 受事/内容/领属物等-事件关系（由在事件前标注\$体现,如"患者""雇员"）
- (j) 工具-事件关系（由在事件前标注*体现，如"手表""计算机"）
- (k) 场所-事件关系（由在事件前标注@体现，如"银行""医院"）
- (l) 时间-事件关系（由在事件前标注@体现，如"假日""孕期"）
- (m) 值-属性关系（直接标注无须借助标识符，如"蓝""慢"）
- (n) 实体-值关系（直接标注无须借助标识符，如"矮子""傻瓜"）
- (o) 事件-角色关系（由加角色名体现，如"购物""盗墓"）
- (p) 相关关系（由在相关概念前标注#体现，如"谷物""煤田"）

3.4 Thesaurus及term自动关联

义原类别

2089

实体 Entity

152

万物 thing (physical, mental, fact)

部分 component (part, fitting)

时间 time

空间 space (direction, location)

事件 Event (relation, state; action)

805

属性 Attribute

245

属性值 AttributeValue

887

次要特征 Secondary feature

128

3.4 Thesaurus及term自动关联

义原的描述

{human|人} {AnimalHuman|动物:

HostOf={Name|姓名}{Wisdom|智慧}

{Ability|能力},{think|思考:agent={~}},

{speak|说:agent={~}}}

3.4 Thesaurus及term自动关联

NO.=030533

W_C=大夫

G_C=noun [da4 fu1]

S_C=

E_C=妇科~, 请~看病, 全国有名的~, 老~, 不会看病的~

W_E=doc

G_E=noun

S_E=

E_E=

DEF={human|人:HostOf={Occupation|职位},

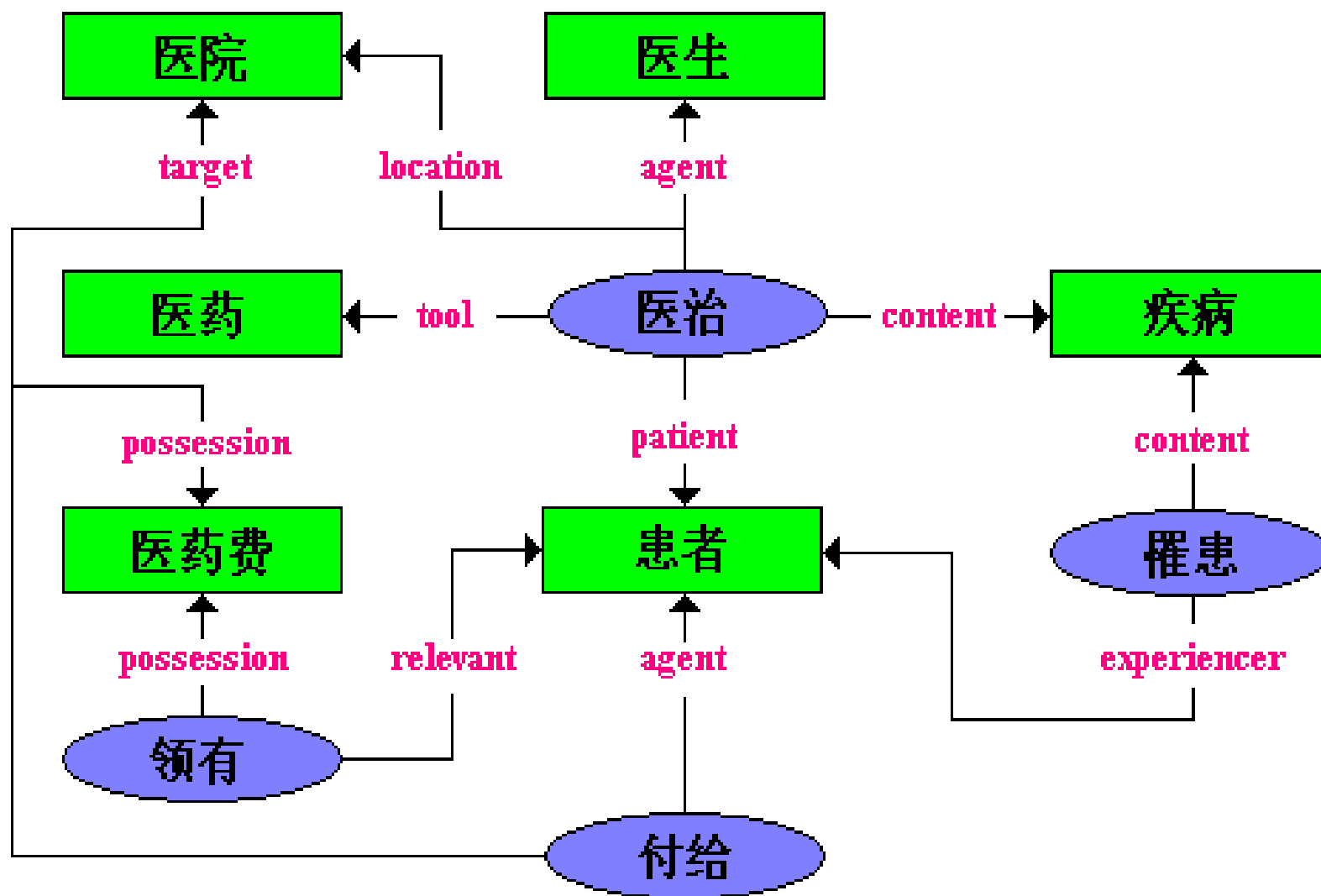
domain={medical|医},{doctor|医治:agent={~}}}

3.4 Thesaurus及term自动关联

概念相似度计算

参赞	皮肤	0.018605
参赞	皮鞋	0.021053
参赞	太太	0.375000
参赞	经理	0.581818
参赞	大使	0.950000

3.4 Thesaurus及term自动关联



3.4 Thesaurus及term自动关联

(2) 梅家驹教授等的《同义词词林》

上海辞书出版社，1983年

12大类，94中类，1428小类

AK02 勇士 侠客 ...

勇士 武士 武夫 壮士 死士 锐士 豪男 虎贲 飞将军

好样的

侠客 侠 武侠 游侠 剑侠 豪侠 豪客 豪士 义士



“武侠” Google 搜索

获得约 32,500,000 条结果 (用时 0.08 秒) 高级搜索

- 所有结果
- 视频
- 新闻
- 博客
- 图书
- 更多

- 网页
- 所有中文网页
- 简体中文网页
- 时间不限
- 最新结果
- 一天内
- 普通视图
- 时光隧道
- 更多搜索工具

武侠小说--天涯在线书库

金庸非**武侠**文集. “大国者下流” · 相思曲与小说 · 北国初春有所思 · 快乐和庄严 · 历史性的一局棋 · 钱学森夫妇的文章 · 书的“续集” · 谈各国象棋 · 围棋杂谈 ...

www.tianyabook.com/wuxia.htm - 网页快照 - 类似结果

武侠小说网- 纯粹的**武侠**小说在线阅读网站

武侠小说网提供金庸, 古龙, 孙晓, 凤歌, 步非烟, 时未寒, 昆仑, 沧海, 英雄志, 等**武侠**小说在线阅读.

昆仑 - 倾城之恋 - 马驹 - 七夜雪

www.wuxia.net.cn/ - 网页快照 - 类似结果

武侠小说-**武侠**小说在线阅读- 铁血读书

铁血读书为您提供各种经典的**武侠**小说、**武侠**小说下载、**武侠**小说在线阅读、**武侠**小说连载, 古典**武侠**、现代**武侠**, 尽在铁血读书。

book.tiexue.net/NovelsWarehouse_1_0_5.html - 网页快照 - 类似结果

武侠电视剧:最新**武侠**片、经典**武侠**片大全- PPTV电视剧

武侠电视剧大全,由PPTV电视剧频道提供. ... 你现在的位置: >> PPTV >> 电视剧 > 电视剧列表 > **武侠**电视剧. 电视剧索引. 按电视剧类型: 科幻 情景剧 历史 励志 警匪 ...

bk.pplive.com/tv/list/class/982 - 网页快照 - 类似结果

武侠 百度百科



“侠客” Google 搜索

获得约 8,260,000 条结果 (用时 0.10 秒)

高级搜索

- 所有结果
- 新闻
- 图片
- 更多

- 网页
- 所有中文网页
- 简体中文网页
- 时间不限
- 最新结果
- 2天内
- 更多搜索工具

相关搜索: [侠客seo](#) [游侠](#) [侠客行电视剧](#)

[侠客手机网手机游戏](#)|[待机图片](#)|[铃声](#)|[视频](#)|[主题](#)|[索爱](#)|[诺基亚](#)|[三星](#) ...

[侠客手机网](#)- Discuz! Board. ... [侠客网](#)个人空间. 看世界. 通过[侠客](#)看世界. [诺基亚](#)官方网站 · [索尼爱立信](#)官方网站 · [Android](#)官方网站 · [三星](#)官方网站 · [摩托罗拉](#)官方 ...
[电子书专区](#) - [索尼爱立信\[SonyEricsson\]大区](#) - [索尼爱立信主题](#) - [空间](#)
[www.mobibal.com/](#) - [网页快照](#) - [类似结果](#)

[侠客_百度百科](#)

2010年8月24日 ... 郑振铎《论武侠小说》：“于是在他们的幼稚的心理上，乃悬盼着有一类‘超人’的[侠客](#)出来。”李白有诗云：“十步杀一人，千里不留行。 ...
[baike.baidu.com/view/4506.htm](#) - [网页快照](#) - [类似结果](#)

[《侠客行》官方网站](#)

《[侠客行](#)》2010年8月21日不删档内测（有奖注册下载/送IPAD）！韩寒选择的网游，拒绝同质、我酷我行！5000万5173现金券注册即送！八大创新非玩不可！韩寒献唱：追梦人！
[xkx.kongzhong.com/](#) - [网页快照](#)

[侠客站长站- 我们更懂个人站长| xkzzz.com](#) 站长[侠客行](#)！

[侠客](#)站长站因站而精彩，为从事站长行业的人士提供丰富的站长资讯、网站运营、建站经验等资讯服务，且提供丰富的站长源码、站长模板、站长插件等必备资源！！
[www.xkzzz.com/](#) - [网页快照](#) - [类似结果](#)



“游侠”

Google 搜索

获得约 8,010,000 条结果 (用时 0.16 秒)

高级搜索

所有结果

▼ 更多

网页

所有中文网页

简体中文网页

时间不限

2天内

▼ 更多搜索工具

[游侠网_单机游戏_中国单机游戏门户](#)

游侠网为单机游戏玩家提供最新单机游戏业界动态、国内外单机游戏下载、游戏补丁、游戏攻略秘籍、游戏专题等内容。坚守单机阵地，弘扬单机文化！

[www.ali213.net/](#) - 网页快照 - 类似结果

- | | |
|--------------------------|-------------------------|
| 火爆论坛 | 星际争霸2 |
| 补丁 | 排行榜 |
| 资源 | 信长之野望13 |
| 实况足球2011 | 文明5 |

[ali213.net](#)站内的其它相关信息 »

[游侠NETSHOW论坛游戏攻略 | 心得秘籍 | 游戏补丁 | 资源下载 | 游戏汉化 ...](#)

游侠NETSHOW论坛中国单机游戏门户站，拥有最广泛、最全面的游戏资讯、补丁、攻略及超高的人气。 - Discuz! Board.

[game.ali213.net/](#) - 网页快照 - 类似结果

[游侠补丁-游侠网](#)

2010年11月2日 ... 游侠补丁网是游侠旗下的一个单机游戏补丁资源库,提供游戏补丁,补丁下载,中英文补丁,原创补丁,免CD补丁,升级档,作弊码,修改器,存档.

[patch.ali213.net/](#) - 网页快照 - 类似结果

[游侠_百度百科](#)



“武侠小说”

Google 搜索

获得约 8,930,000 条结果 (用时 0.14 秒)

[高级搜索](#)

所有结果

更多

网页

所有中文网页

简体中文网页

时间不限

2天内

更多搜索工具

[武侠小说--天涯在线书库](#)

其他武侠名家作品集: . 还珠楼主小说集 柳残阳小说集 司马翎作品集 司马紫烟作品集 云中岳作品集 诸葛青云作品集 其他更多 **武侠小说**: 共3页第 (1) (2) (3) 页 ...

www.tianyabook.com/wuxia.htm - [网页快照](#) - [类似结果](#)

[武侠小说网- 纯粹的武侠小说在线阅读网站](#)

武侠小说网提供金庸, 古龙, 孙晓, 凤歌, 步非烟, 时未寒, 昆仑, 沧海, 英雄志, 等 **武侠小说** 在线阅读.

www.wuxia.net.cn/ - [网页快照](#) - [类似结果](#)

[武侠小说](#)

东方白小说 · 倪匡 **武侠小说** · 上官鼎小说 · 刘定坚小说 · 天宇小说 · 陈青云小说 · 独孤红小说 · 慕容美小说 · 罗森小说 · 江和小说 · 诸葛青云小说 · 忆文小说 ...

www.shuku.net:8080/novels/mulu/wuxia.html - [网页快照](#)

[《小说阅读网》 - 武侠小说,仙侠小说](#)

《小说阅读网》是 **武侠小说**、仙侠小说最大的原创网站之一, 提供最新热门好看的 **武侠小说**、仙侠小说全文在线免费阅读、经典完结 **武侠小说**、仙侠小说排行榜和 **武侠小说**、仙侠 ...

www.readnovel.com/ch/3.html - [网页快照](#) - [类似结果](#)

[百万书库-->武侠小说](#)

百万书库>**武侠小说**. **武侠小说**. 作品集. 金庸 · 古龙 · 梁羽生 · 温瑞安 · 黄易 · 倪匡 · 萧逸



“侠客小说”

Google 搜索

获得约 7,740 条结果 (用时 0.17 秒)

[高级搜索](#)

所有结果

更多

网页

[所有中文网页](#)

[简体中文网页](#)

更多搜索工具

[从淫贼到侠客-小说书](#)

类别: 武侠修真, 作者: 剑痕泪, 管理员: , 全文长度: 166113字. 最后更新: 2008-07-26, 文章状态: 连载中, 授权级别: 暂未授权, 首发状态: 他站首发 ...

www.xiaoshuoshu.cn/files/article/info/0/98.htm - [网页快照](#)

[侠客小说的科幻梦——评黄易《寻秦记》](#)

从文种的立场而言, 艰深小说象样大抵分成侠客、言情、科幻、历史、侦察五大种, 各有其从成一格的种型特点侠客之侠客、言情之恋情、科幻之科学幻想、历史之历史、侦察之 ...

www.baoshop.info/siwaanmo/233.html - [网页快照](#)

[猎人侠客小说_百度知道](#)

2010年8月24日 ... 西索: 《卖花朵的小女孩》《爱上西索的悲惨情史》《天啊, 我只想活下去啊!》《猎人--下弦之月》《猎人之天意使然》《[猎人]我与西索不得不说的故事》《猎人- ...

zhidao.baidu.com/question/177877334.html?push=related - [网页快照](#)

[【小说】网游之逍遥侠客|网游之逍遥侠客最新章节|全集下载_飞卢小说 ...](#)

本站提供星海浪涛的网游之逍遥**侠客小说**最新章节在线阅读,全文阅读及网游之逍遥侠客全集下载,网游之逍遥侠客txt下载等,希望本站能给您的阅读带来安静与喜悦.

b.faloo.com/f/29970.html - [网页快照](#)

[侠客小说- 标签索引- 红袖添香](#)



网页 图片 视频 地图 新闻 音乐 购物 Gmail 更多 ▼



“游侠小说”

Google 搜索

获得约 13,800 条结果 (用时 0.19 秒)

[高级搜索](#)

所有结果

更多

网页

所有中文网页

简体中文网页

时间不限

3 天内

普通视图

图文并茂

更多搜索工具

[网游之神迹游侠- 小说520](#)

网游之神迹游侠,小说520. ... 【网游之神迹游侠小说简介】. 提剑斩鲸的虚拟网游小说新书——<网游之神迹游侠> 在现实中找不到方向的白羽，能否在虚拟世界中实现他的 ...

www.xiaoshuo520.com/Book/66574/index.aspx - 网页快照

[【小说】黑道游侠|黑道游侠最新章节|全集下载_飞卢小说免费在线阅读](#)

本站提供丁小生的黑道游侠小说在线阅读,全文阅读及黑道游侠全集下载,希望本站能给您的阅读带来安静 ... 已有504人读过黑道游侠小说已写45618字... 目前仍在拼命写作中. ...

b.faloo.com/f/84880.html - 网页快照

[颜倾天下前传神曲倦爱_碧游侠小说_迅雷小说下载_TXT电子书迅雷下载_书 ...](#)

2009年10月3日 ... 颜倾天下前传神曲倦爱_碧游侠小说_迅雷小说下载_TXT电子书迅雷下载_书友小说网.

suucn.com/html/2009103125/112623/ - 网页快照

[异界之黑暗游侠|首发小说异界之黑暗游侠最新章节- 123读小说网](#)

异界之黑暗游侠小说专题简介：. 123读123读像首歌，让我们用整齐的声音唱出123读！本站为您提供小说异界之黑暗游侠全集章节内容，坚决免费阅读，这是书友看异界之黑暗 ...

www.123du.net/dudu-32/355769/ - 网页快照

[碧游侠小说作品集](#)

碧游侠小说作品集



“武侠行”

Google 搜索

获得约 5,410 条结果 (用时 0.19 秒)

高级搜索

所有结果

视频

更多

网页

所有中文网页

简体中文网页

更多搜索工具

[\[武侠\]行镖记 仗剑天涯 天涯社区](#)

[武侠]行镖记. 点击: 606 回复: 44. 作者: 普祥真人 发表日期: 2010-4-12 21:10:00. 清康熙五十六年, 夏五月。 五月里的北京城, 骄阳似火, 那太阳将甬路都仿佛要晒化了 ...

www.tianya.cn/publicforum/content/no17/1/44442.shtml - 网页快照

有关“[“武侠行”](#)”的视频



[武侠行之武侠成功经](#)

4 分钟 - 2008年11月7日

[v.ku6.com](#)



[武侠行之武侠成功经](#)

4 分钟 - 2009年6月2日

[v.youku.com](#)

[青武侠行的动态 i贴吧](#)

查看青[武侠行](#)的过往发言>>. 关于青[武侠行](#). 简介: 大家好, 我是青[武侠行](#), 欢迎来到我的i贴吧! 你可以通过关注我, 来了解我的i贴吧动向... 贴吧豆消费: 0个 贴吧商城 ...

tieba.baidu.com/i/123000552?st_mod=pb&fr=tb0...st... - 网页快照

[索爱W580i武侠行 手机游戏下载 中国手机游戏中心\[game1313.com\]](#)

2010年10月22日 ... 中国手机游戏中心, 提供免费[武侠行](#)手机游戏下载, 更多免费索爱W580i [武侠行](#)手机游戏下载.

www.game1313.com/W580i/v29275.html - 网页快照

[诺基亚N79武侠行 手机游戏下载 中国手机游戏中心\[game1313.com\]](#)

中国手机游戏中心, 提供免费[武侠行](#)手机游戏下载, 更多免费诺基亚N79[武侠行](#)手机游戏 ...

[1313.com/W580i/v29275.html">1313.com/W580i/v29275.html](#) - 网页快照



“侠客行”

Google 搜索

获得约 989,000 条结果 (用时 0.08 秒)

高级搜索

所有结果

视频

新闻

更多

网页

所有中文网页

简体中文网页

时间不限

最新结果

2天内

更多搜索工具

[《侠客行》官方网站](#)

《侠客行》2010年8月21日不删档内测（有奖注册下载/送IPAD）！韩寒选择的网游，拒绝同质、我酷我行！5000万5173现金券注册即送！八大创新非玩不可！韩寒献唱：追梦人！

xkx.kongzhong.com/ - 网页快照

[侠客行（新版全集） - 专辑- 优酷视频](#)

相关专辑. [侠客行](#); 播放: 12822. [侠客行](#) (梁朝伟); 播放: 15423. 全集; 播放: 2613. 咏春(全集); 播放: 109234. 蓝狐 (全集); 播放: 57126. 短刀行; 播放: 13494 ...

www.youku.com > 专辑 > 电视剧 - 网页快照 - 类似结果

[侠客行_百度百科](#)

2010年6月7日 ... 《侠客行》原是一首描写和歌颂侠客的古体五言诗，由李白所作。金庸的一部武侠小说《侠客行》，初次发表为1965年，据说灵感来自李白的“古风五十九首” ...

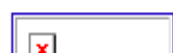
baike.baidu.com/view/4383.htm - 网页快照 - 类似结果

[白鹿书院----侠客行](#)

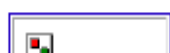
第13章舐犊之情 · 第14章关东四大门派 · 第15章真相 · 第16章凌霄城 · 第17章自大成狂 · 第18章有所求 · 第19章腊八粥 · 第20章“侠客行” · 第21章“我是谁?” ...

www.oklink.net/wxsj/jing-yong/knight.../index.html - 网页快照 - 类似结果

[有关“《侠客行》”的视频](#)



[侠客行38](#)



[侠客行](#)




“游侠行”

Google 搜索

获得约 20,800 条结果 (用时 0.18 秒)

[高级搜索](#)

 所有结果

 更多

网页

所有中文网页

简体中文网页

 更多搜索工具

[杂曲歌辞·游侠行 赏析孟郊 杂曲歌辞·游侠行 鉴赏 杂曲歌辞·游侠行 的 ...](#)

唐 时代孟郊 的杂曲歌辞·[游侠行](#) 赏析,杂曲歌辞·[游侠行](#) 的诗意、杂曲歌辞·[游侠行](#) 诗歌全文、杂曲歌辞·[游侠行](#) 名句.

www.jlonline.com/shici/shige23958/ - [网页快照](#)

[游侠行摄的旅游空间| Yododo 游多多](#)

2010年7月8日 ... [游侠行](#)摄的旅游空间: 乐山乐水用心用情感受自然的美景.

www.yododo.com/.../0129AFDB44280B2DFF80808129AE8145 - [网页快照](#)

[漫步在自由道- 天津游记- 游侠行摄的游记| Yododo 游多多](#)

天津游记, 昨天晚上道天津已经是深夜了, 来过天津好几次却从没有住下过, 这次因为学 ...

www.yododo.com/.../012BE5E89AA50503FF8080812BE4FD23 - [网页快照](#)

 [显示来自 yododo.com 的更多搜索结果](#)

[诗.词.意- 唐五代.孟郊.游侠行\(壮士性刚决\)](#)

诗.词.意, 提供中国文学, 诗词查询的网站. 提供文学交流的地方和平台。

www.rdlou.com/poem/index.asp?scid=12046 - [网页快照](#)

[轻松订门票游侠行天下 - 旅游名店城](#)

轻松订门票[游侠行](#)天下一一2009GIFT银旅通活动面面观. 2009年3月26日~29日银旅通联合中国银联广东分公司、韶关旅游局、肇庆旅游发展局、工商银行广东省分公司、中国移动 ...

www.yocity.cn/yocity_effect.asp?id=5365 - [网页快照](#)

3.5 Thesaurus及term自动关联

Problems with resources like WordNet

- Great as a resource but missing nuance
 - e.g., “proficient” is listed as a synonym for “good”
This is only correct in some contexts
- Missing new meanings of words
 - e.g., wicked, badass, nifty, wizard, genius, ninja, bombest
 - Impossible to keep up-to-date!
- Subjective
- Requires human labor to create and adapt
- Can't compute accurate word similarity →

3.5 Thesaurus及term自动关联

- Attempt to generate a thesaurus automatically by analyzing the collection of documents
- Fundamental notion: similarity between two words
 - Automatic thesaurus construction by contexts
- **Definition: Two words are similar if they co-occur with similar words.**
- You can harvest, peel, eat, prepare, etc. apples and pears, so apples and pears must be similar.

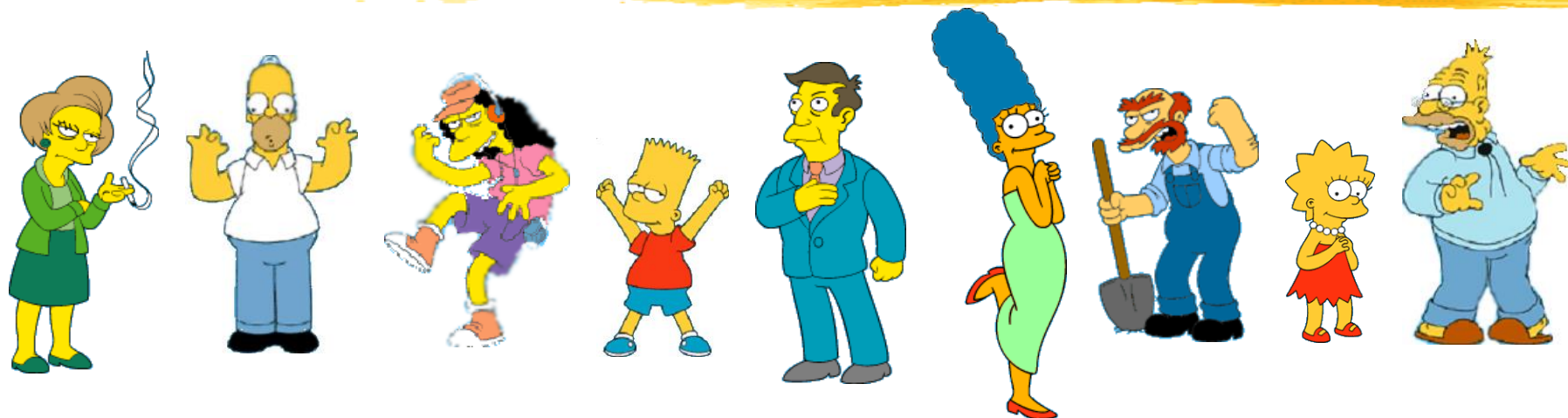
3.5 Thesaurus及term自动关联



● What is Clustering

- * Partition a set of objects into groups or clusters.
 - Similar objects are placed in the same group and dissimilar objects in different groups.
 - High intra-class similarity
 - Low inter-class similarity
- * Objects are described and clustered using a set of features and values.

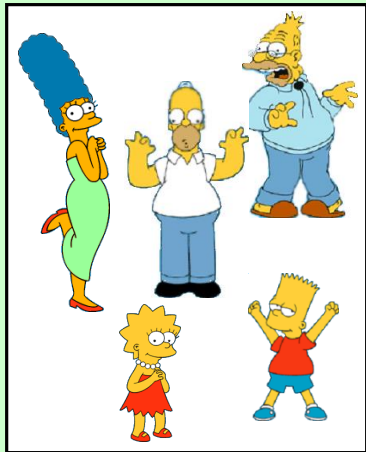
3.5 Thesaurus及term自动关联



3.5 Thesaurus及term自动关联



Clustering is subjective



Simpson's Family



School Employees



Females



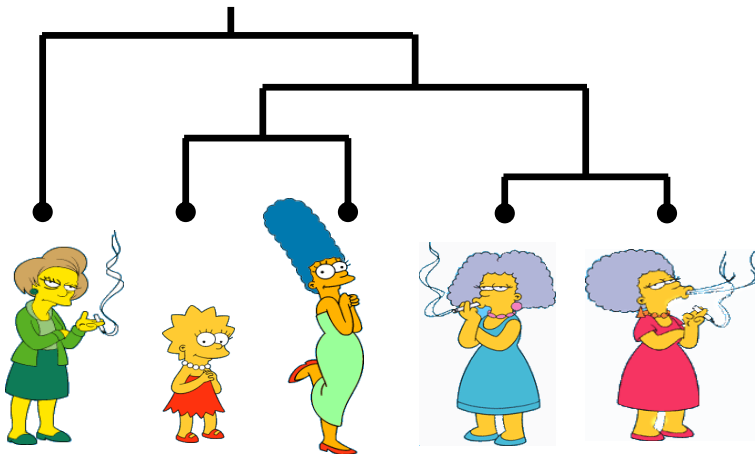
Males

3.5 Thesaurus及term自动关联

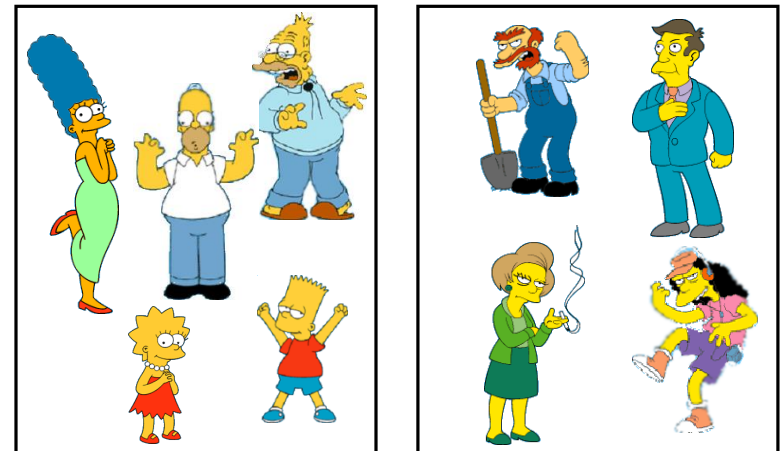
Two Types of Clustering

- * Partitional clustering (Non-hierarchical)
 - Construct various partitions and then evaluate them by some criterion
- * Hierarchical clustering
 - Create a hierarchical decomposition of the set of objects using some criterion

Hierarchical



Partitional



3.5 Thesaurus及term自动关联

Partitional Clustering

K-means clustering

1. Decide on a value for k .
2. Initialize the k cluster centers (randomly, if necessary).
3. Decide the class memberships of the N objects by assigning them to the nearest cluster center.
4. Re-estimate the k cluster centers, by assuming the memberships found above are correct.
5. If none of the N objects changed membership in the last iteration, exit. Otherwise goto 3.

3.5 Thesaurus及term自动关联

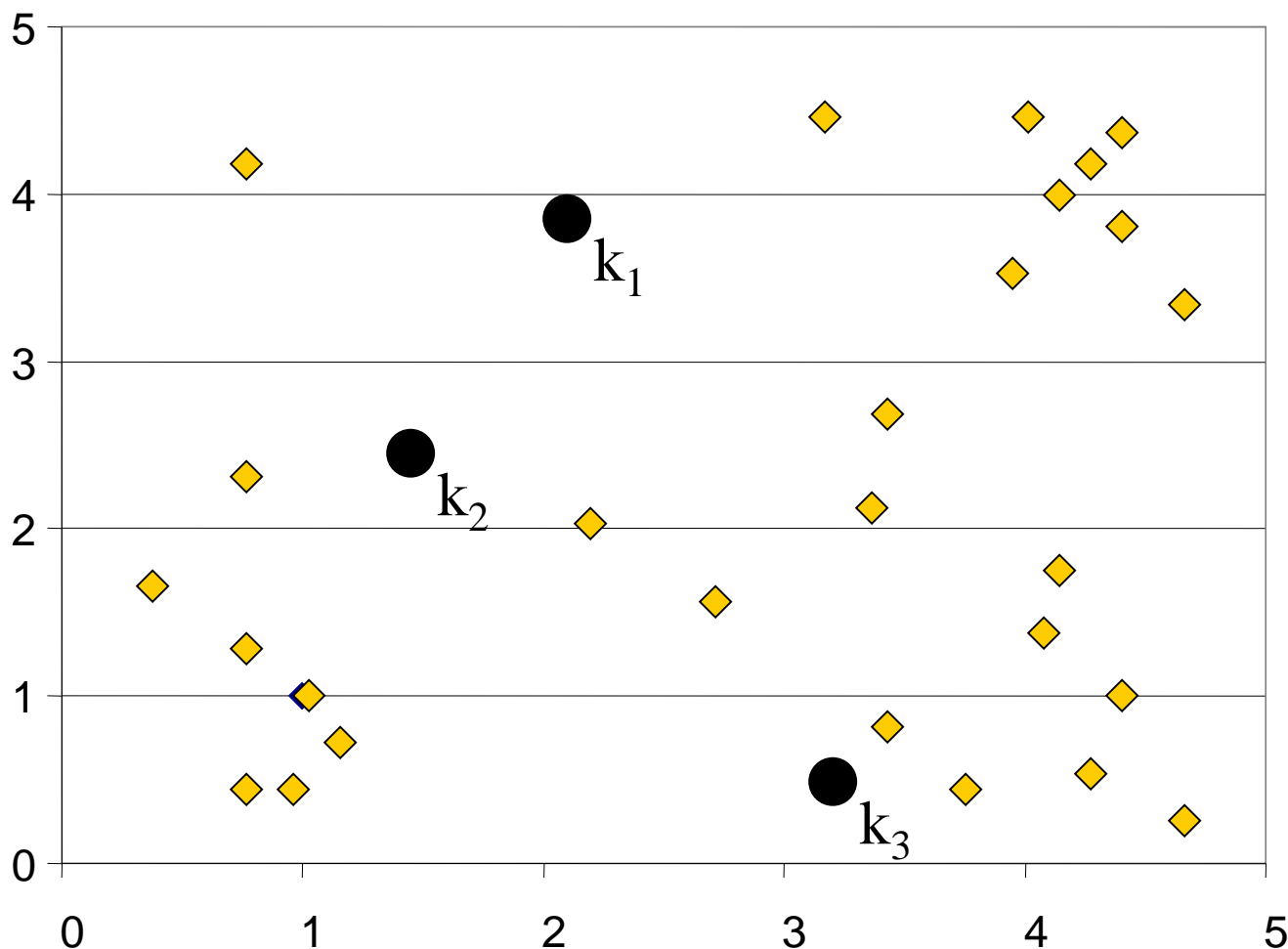
Key notion: cluster centroid

Centroid of a cluster = average of vectors
in a cluster - is a vector.



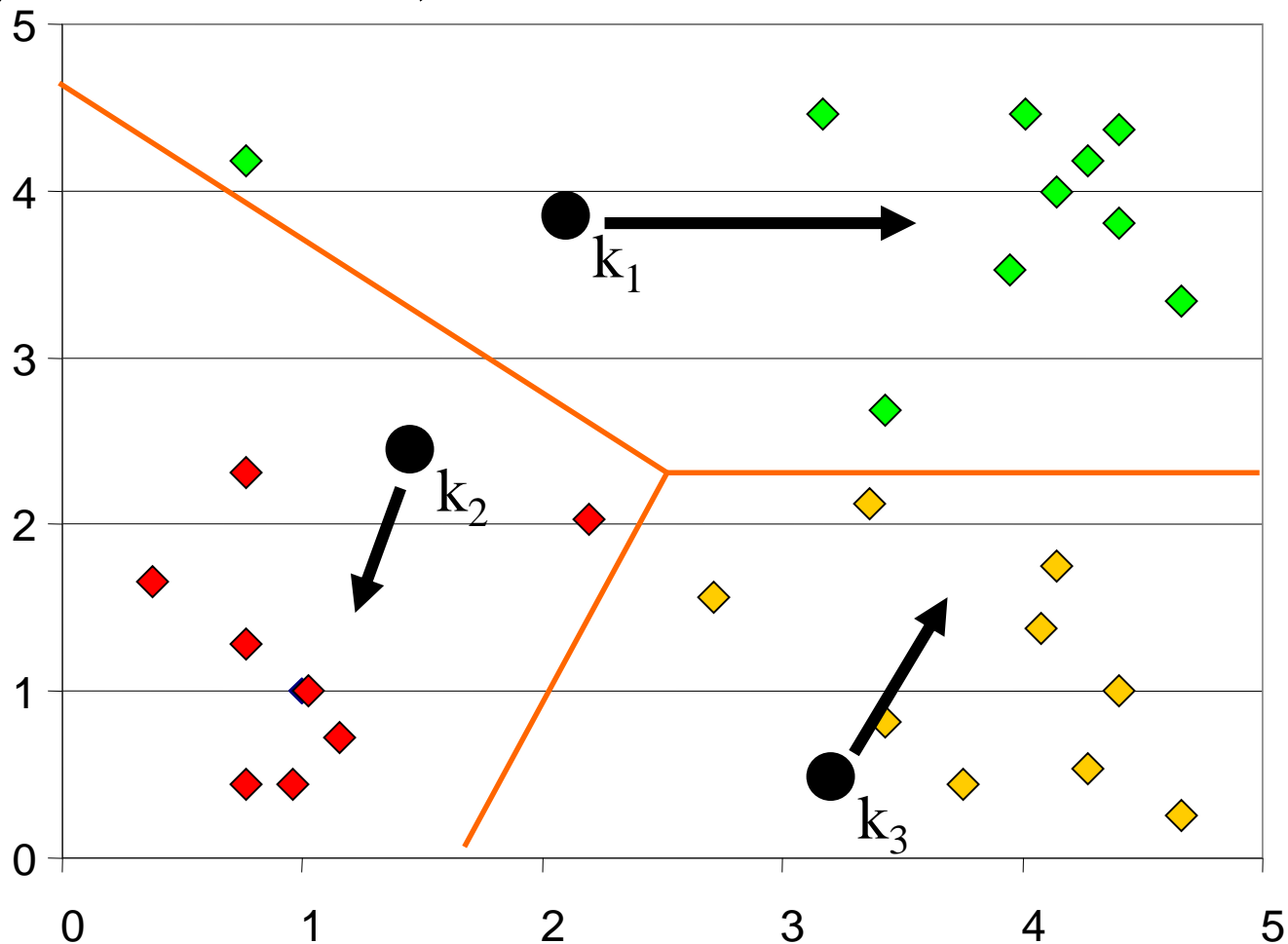
3.5 Thesaurus及term自动关联

Algorithm: k-means, Distance Metric: Euclidean Distance



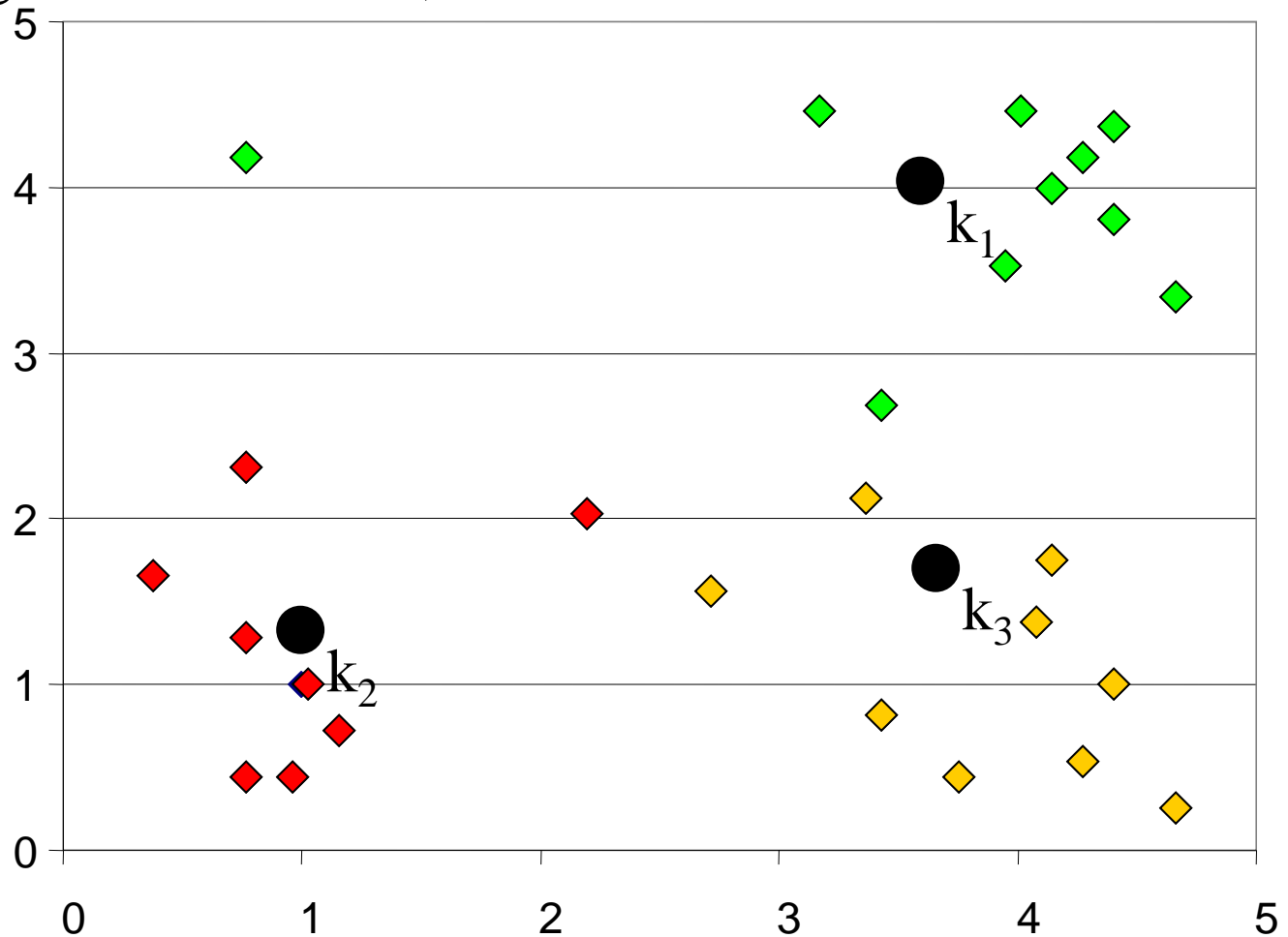
3.5 Thesaurus及term自动关联

Algorithm: k-means, Distance Metric: Euclidean Distance



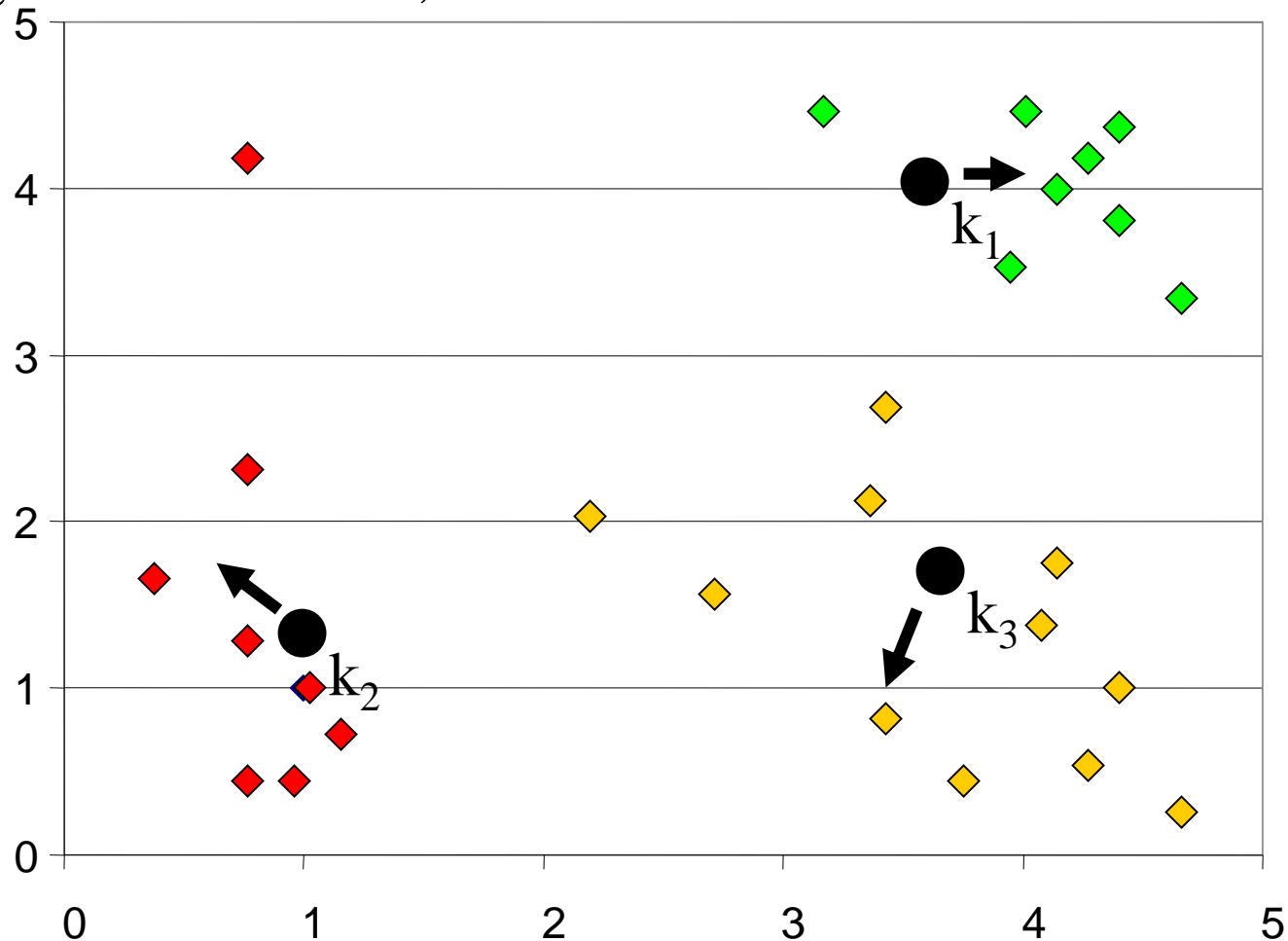
3.5 Thesaurus及term自动关联

Algorithm: k-means, Distance Metric: Euclidean Distance



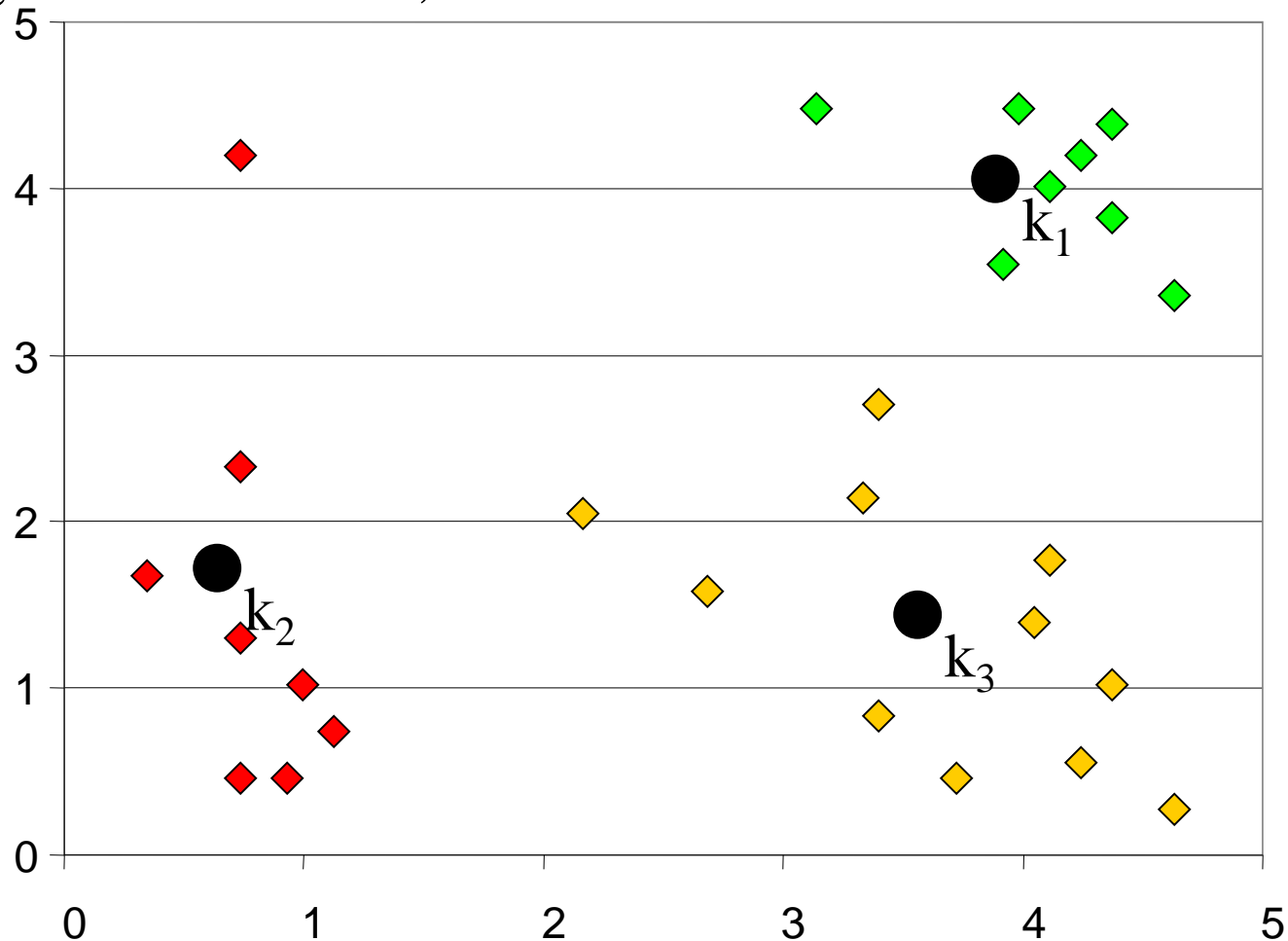
3.5 Thesaurus及term自动关联

Algorithm: k-means, Distance Metric: Euclidean Distance



3.5 Thesaurus及term自动关联

Algorithm: k-means, Distance Metric: Euclidean Distance



3.5 Thesaurus及term自动关联



Termination conditions

Several possibilities, e.g.,

- A fixed number of iterations.

- Term partition unchanged.

- Centroid positions don't change.

3.5 Thesaurus及term自动关联

Comments on the K-means method

* Advantages

- *Relatively efficient: $O(tkn)$* , where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t \ll n$.
- Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* and *genetic algorithms*

* Disadvantages

- Applicable only when *mean* is defined, then what about categorical data?
- Need to specify k , the *number* of clusters, in advance
- Unable to handle noisy data and *outliers*
- Not suitable to discover clusters with *non-convex shapes*

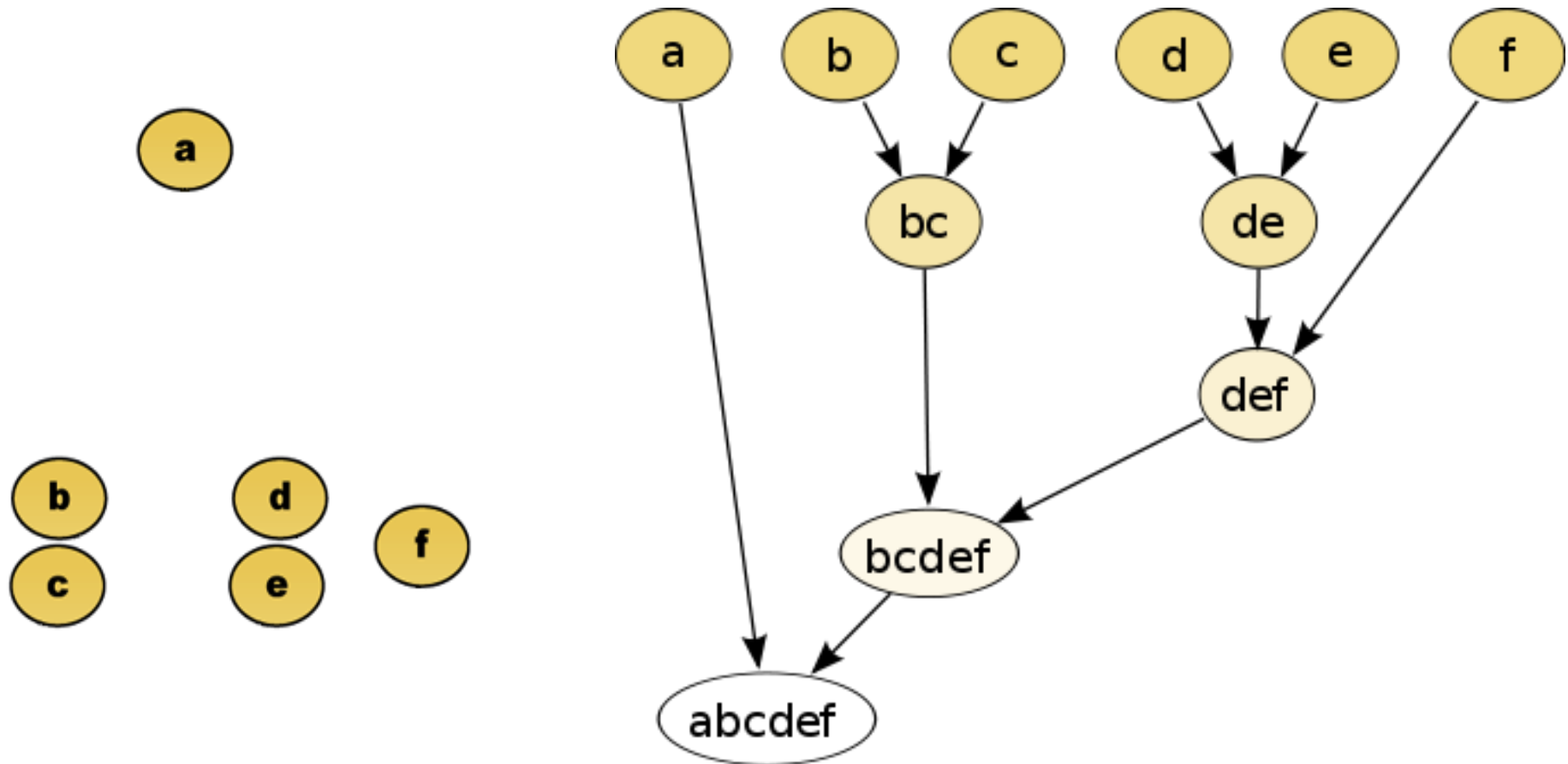
3.5 Thesaurus及term自动关联



Hierarchical Clustering


- * **Bottom-Up (agglomerative):** Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.
- * **Top-Down (divisive):** Starting with all the data in a single cluster, consider every possible way to divide the cluster into two. Choose the best division and recursively operate on both sides.

3.5 Thesaurus及term自动关联

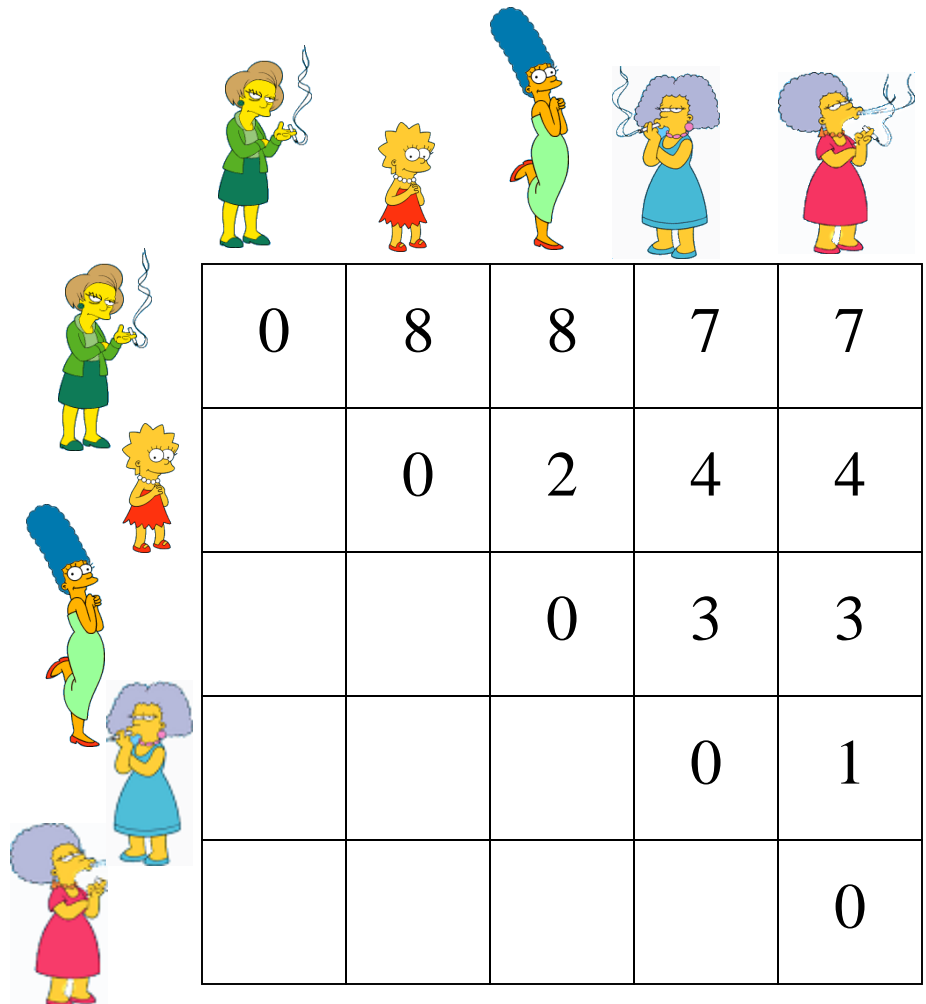


3.5 Thesaurus及term自动关联











We begin with a distance matrix which contains the distances between every pair of objects in our database.


$$D(\text{Mrs. Simpson}, \text{Lisa Simpson}) = 8$$


$$D(\text{Marge Simpson}, \text{Lisa Simpson}) = 1$$

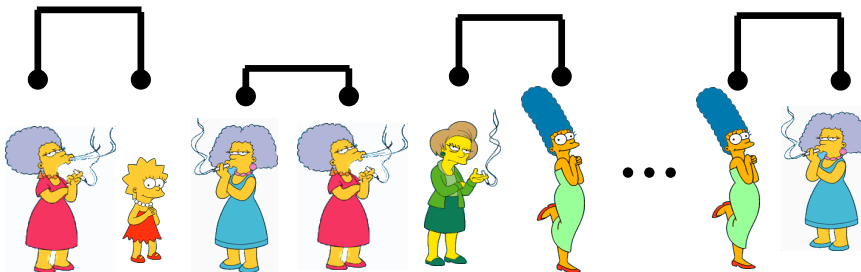


A diagram showing the Simpson family members (Mrs. Simpson, Lisa Simpson, Marge Simpson, Marge Simpson, and Lisa Simpson) arranged in a grid, with a distance matrix table to their right. The table contains numerical values representing distances between the objects.

					
	0	8	8	7	7
		0	2	4	4
			0	3	3
				0	1
					0

Bottom-Up: Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

Consider all possible merges...

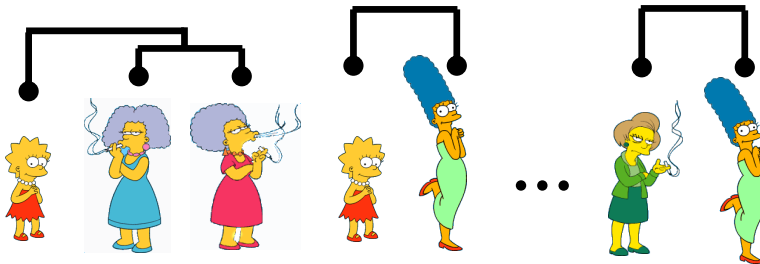


Choose the best

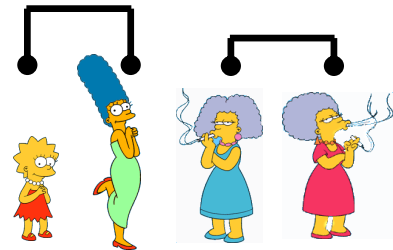


Bottom-Up: Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

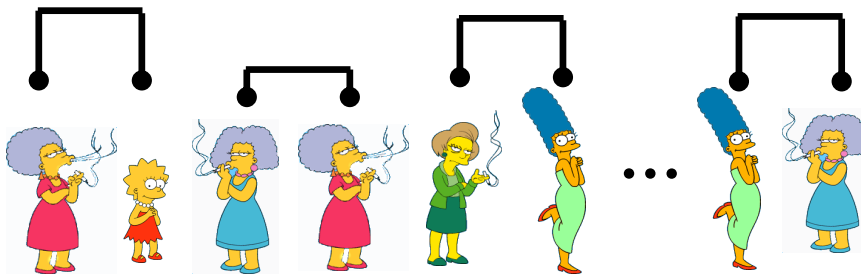
Consider all possible merges...



Choose the best



Consider all possible merges...

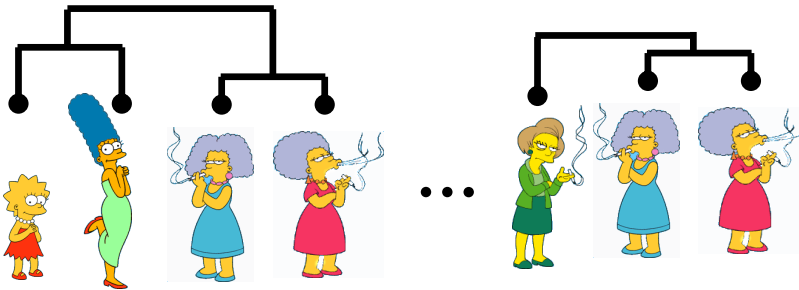


Choose the best

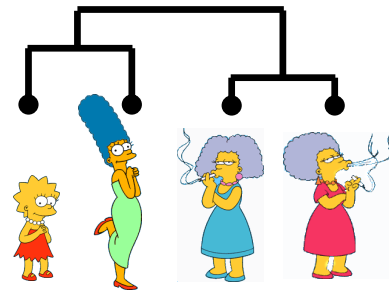


Bottom-Up: Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

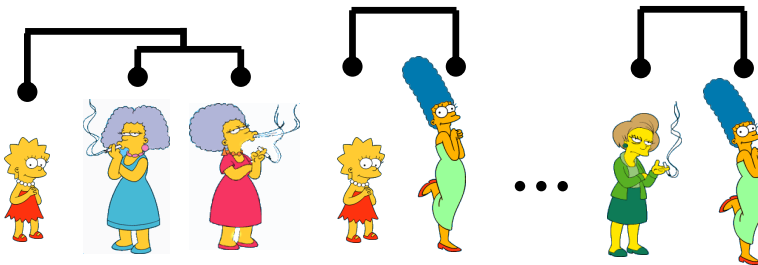
Consider all possible merges...



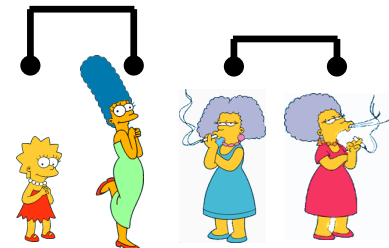
Choose the best



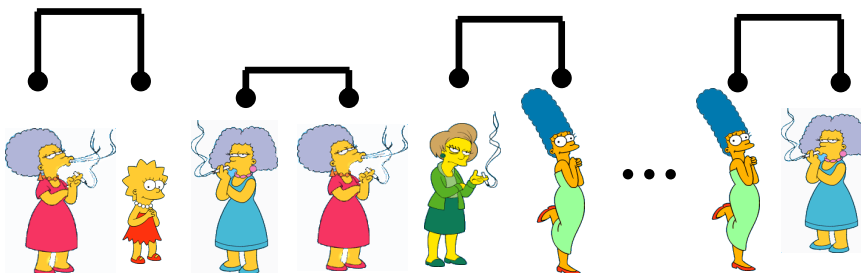
Consider all possible merges...



Choose the best



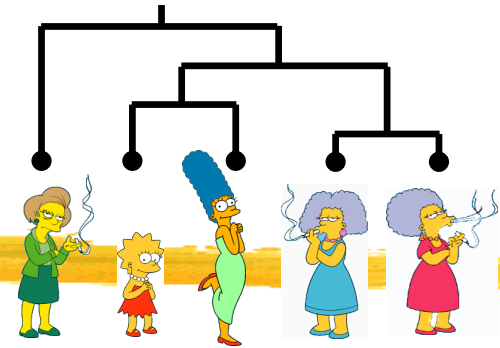
Consider all possible merges...



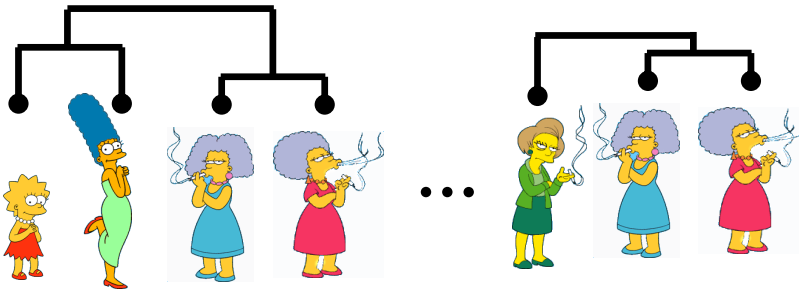
Choose the best



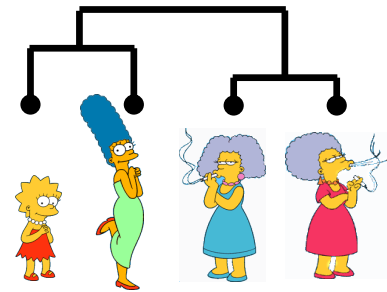
Bottom-Up: Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



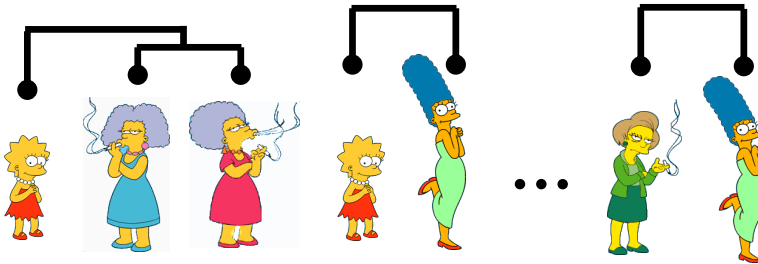
Consider all possible merges...



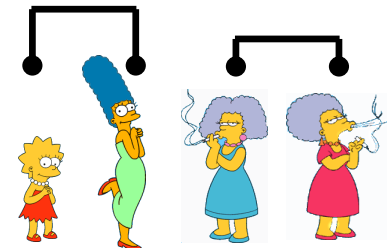
Choose the best



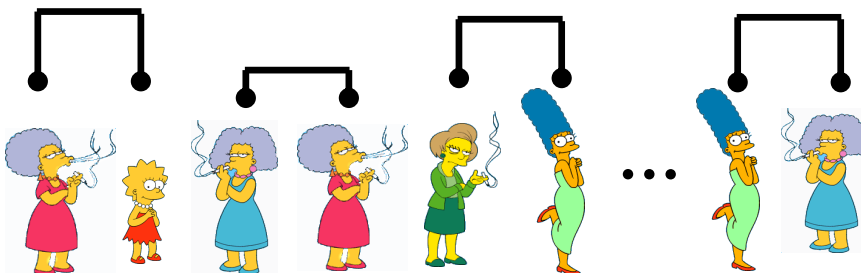
Consider all possible merges...



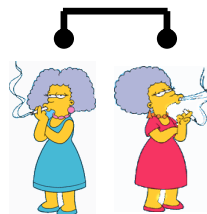
Choose the best



Consider all possible merges...

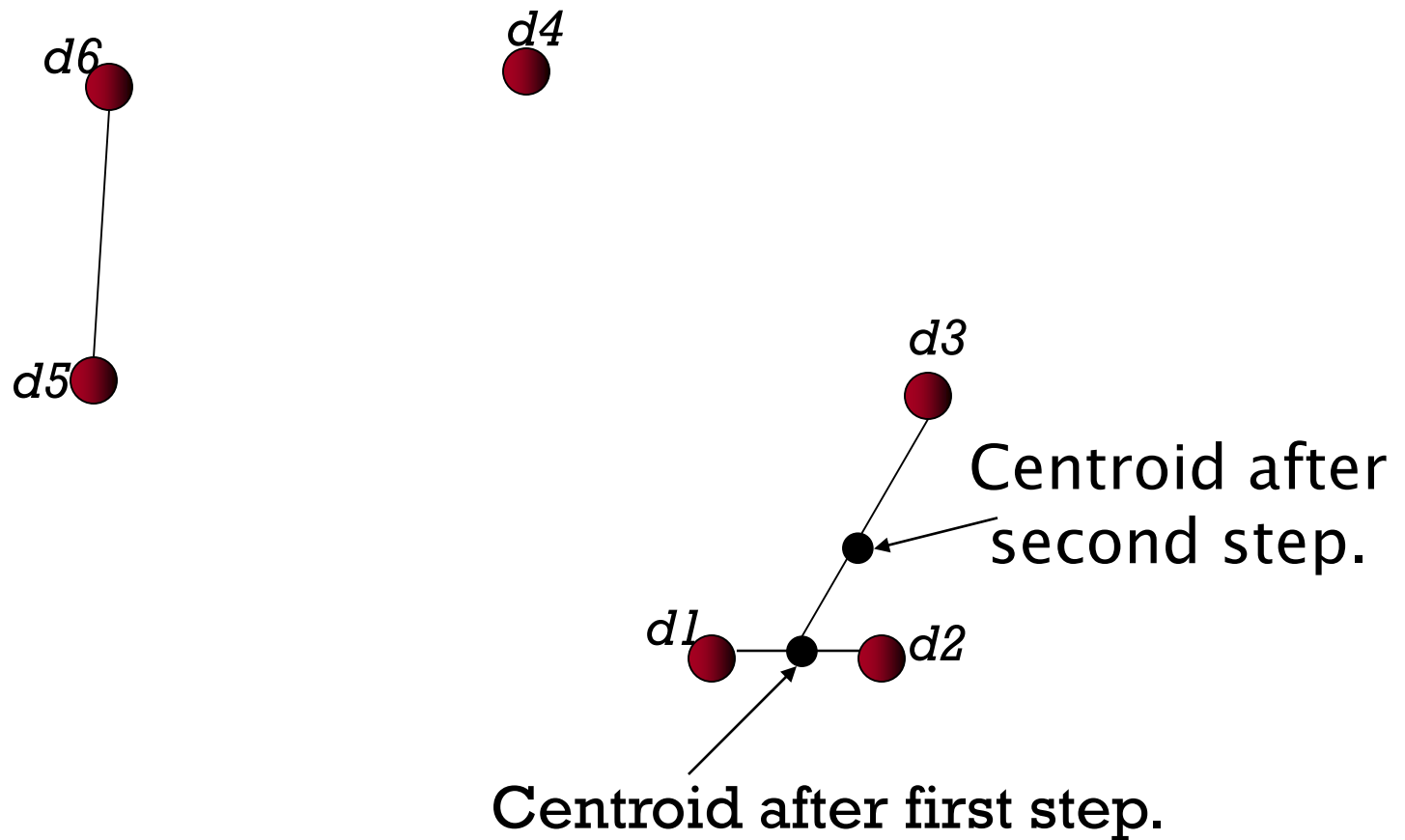


Choose the best



3.5 Thesaurus及term自动关联

Example: $n=6$, $k=3$, closest pair of centroids



3.5 Thesaurus及term自动关联

Similarity measure

- * **Single linkage (nearest neighbor):** In this method the distance between two clusters is determined by the distance of the two closest objects (nearest neighbors) in the different clusters.
- * **Complete linkage (furthest neighbor):** In this method, the distances between clusters are determined by the greatest distance between any two objects in the different clusters (i.e., by the "furthest neighbors").
- * **Group average linkage:** In this method, the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters.

3.5 Thesaurus及term自动关联

“closest pair” of clusters:

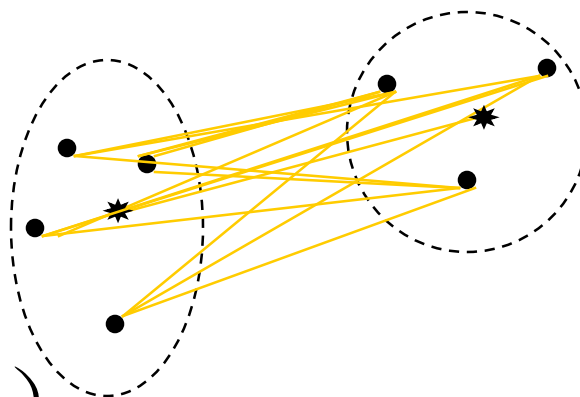
涉及类间距离的定义

最近距离

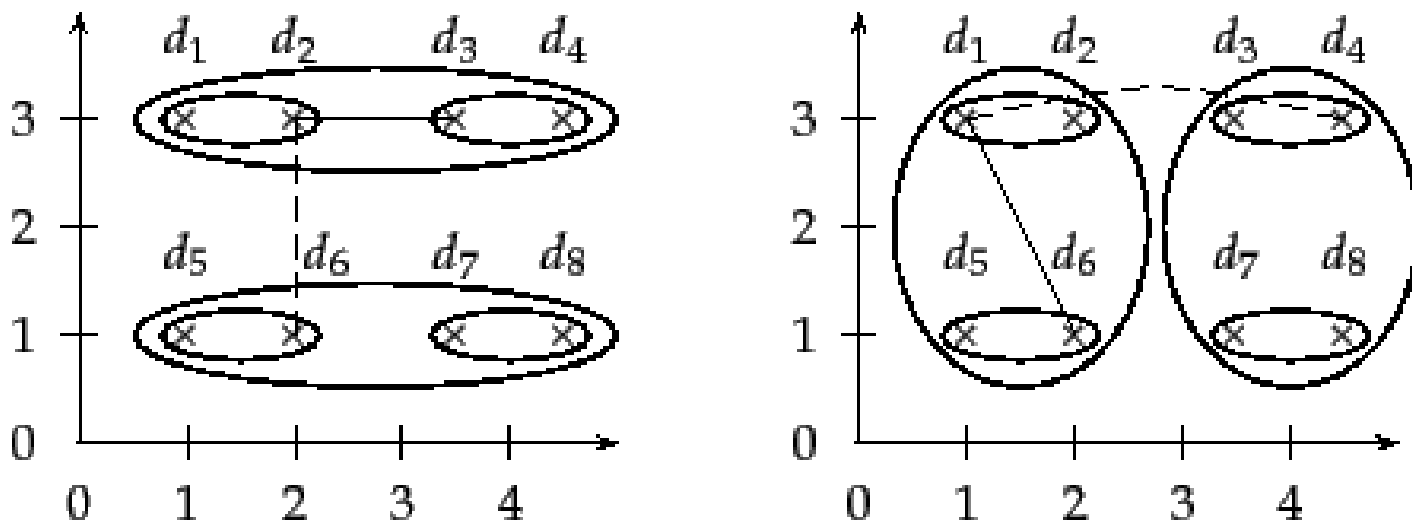
最远距离

均值距离

(centroid
之间的距离)

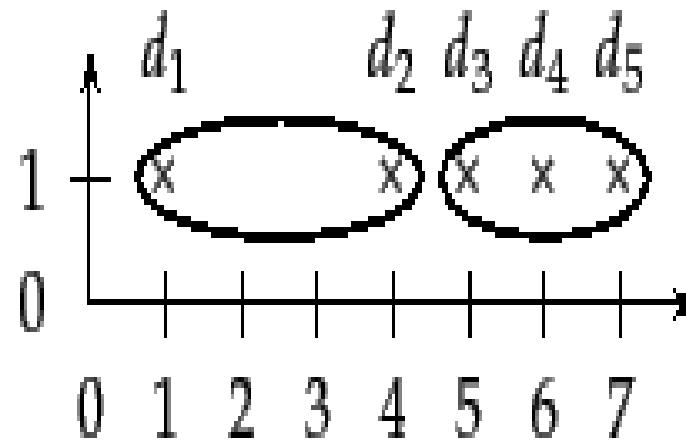


3.5 Thesaurus及term自动关联

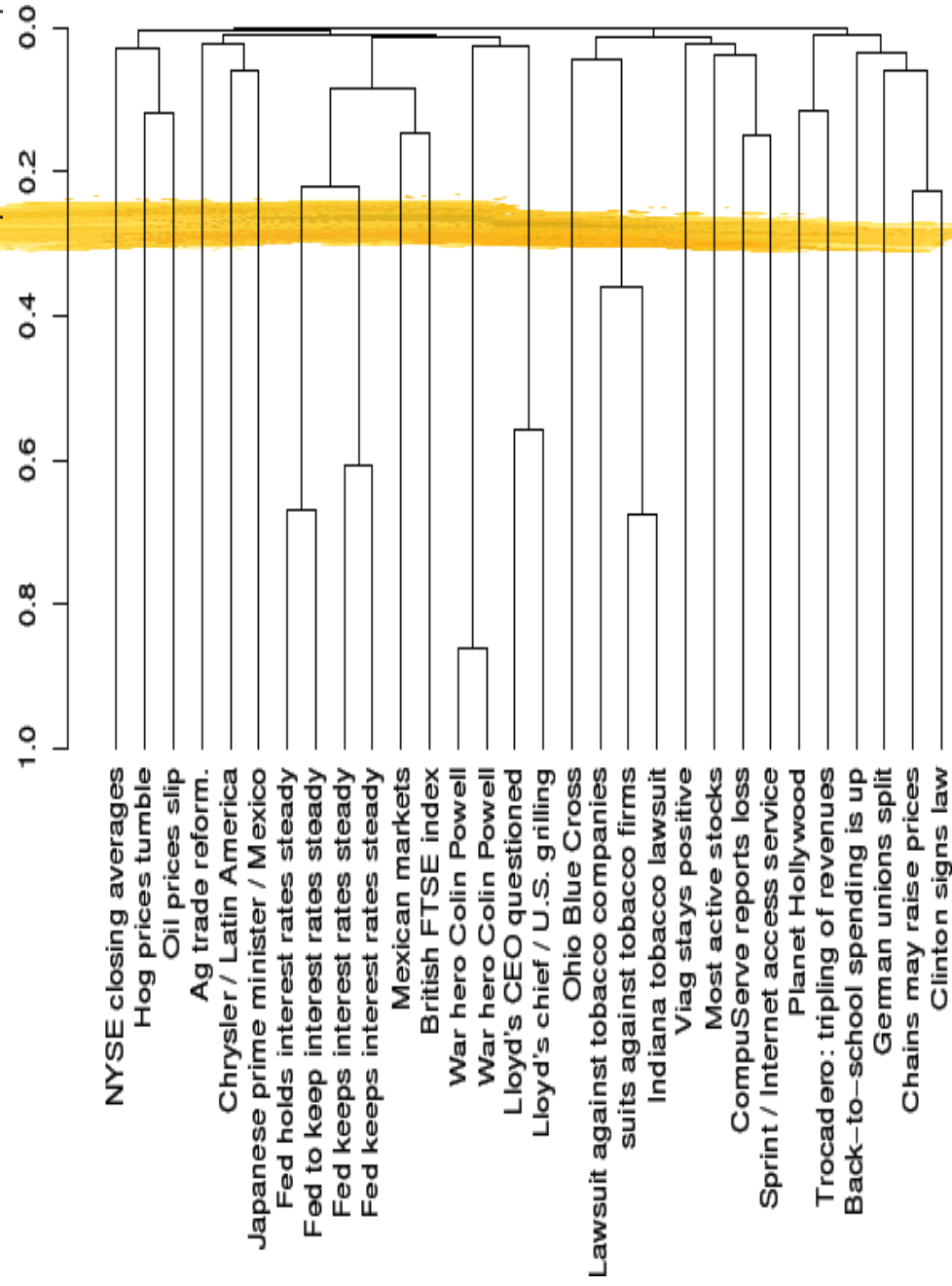
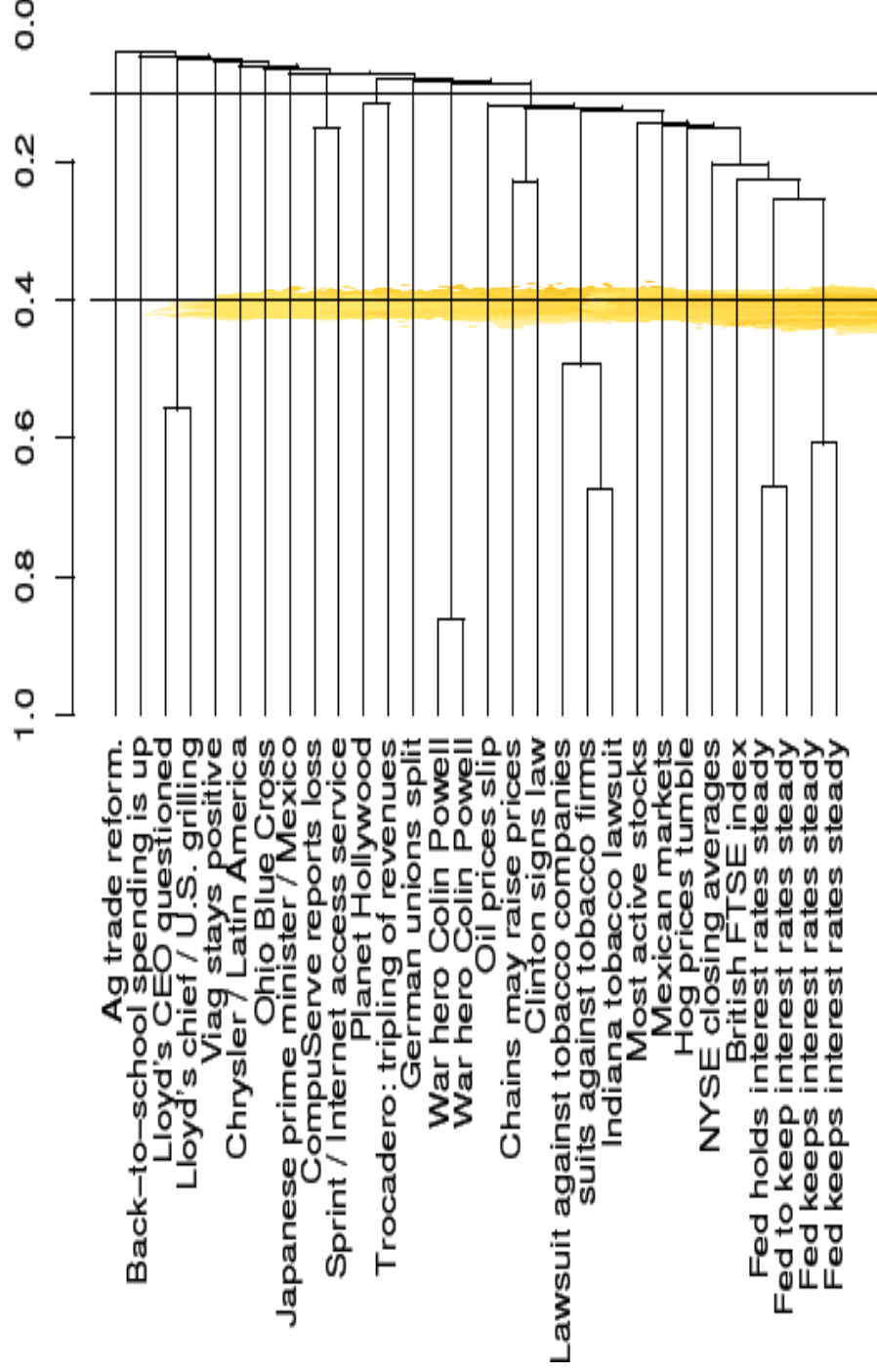


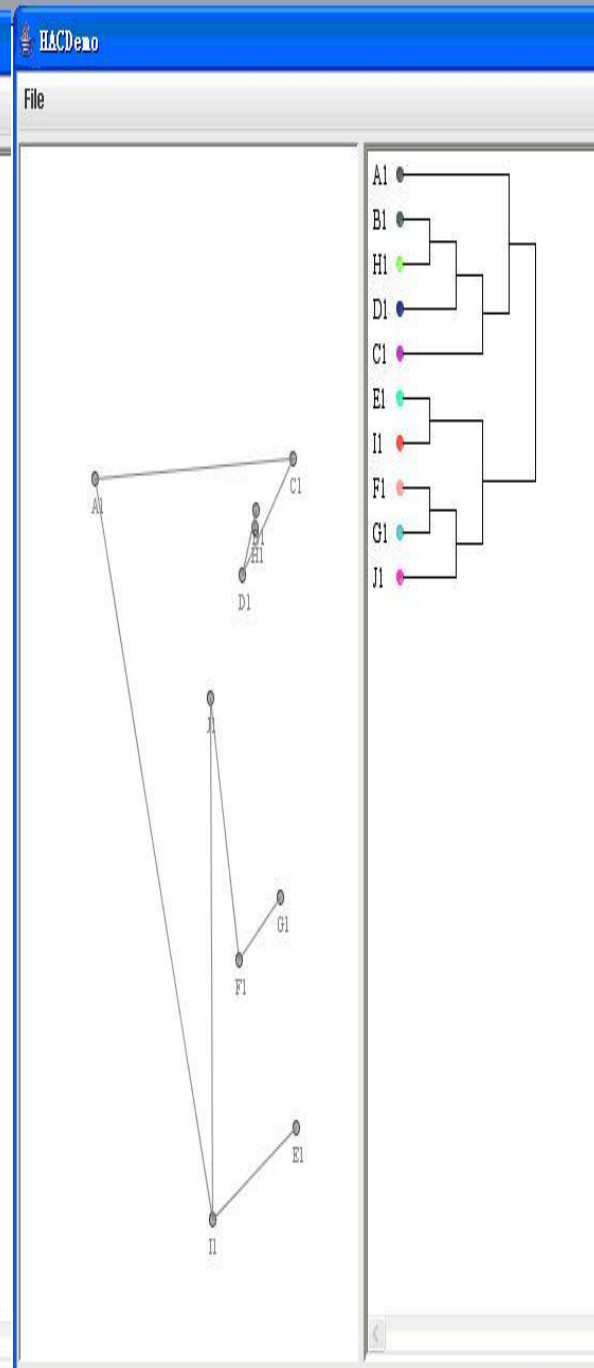
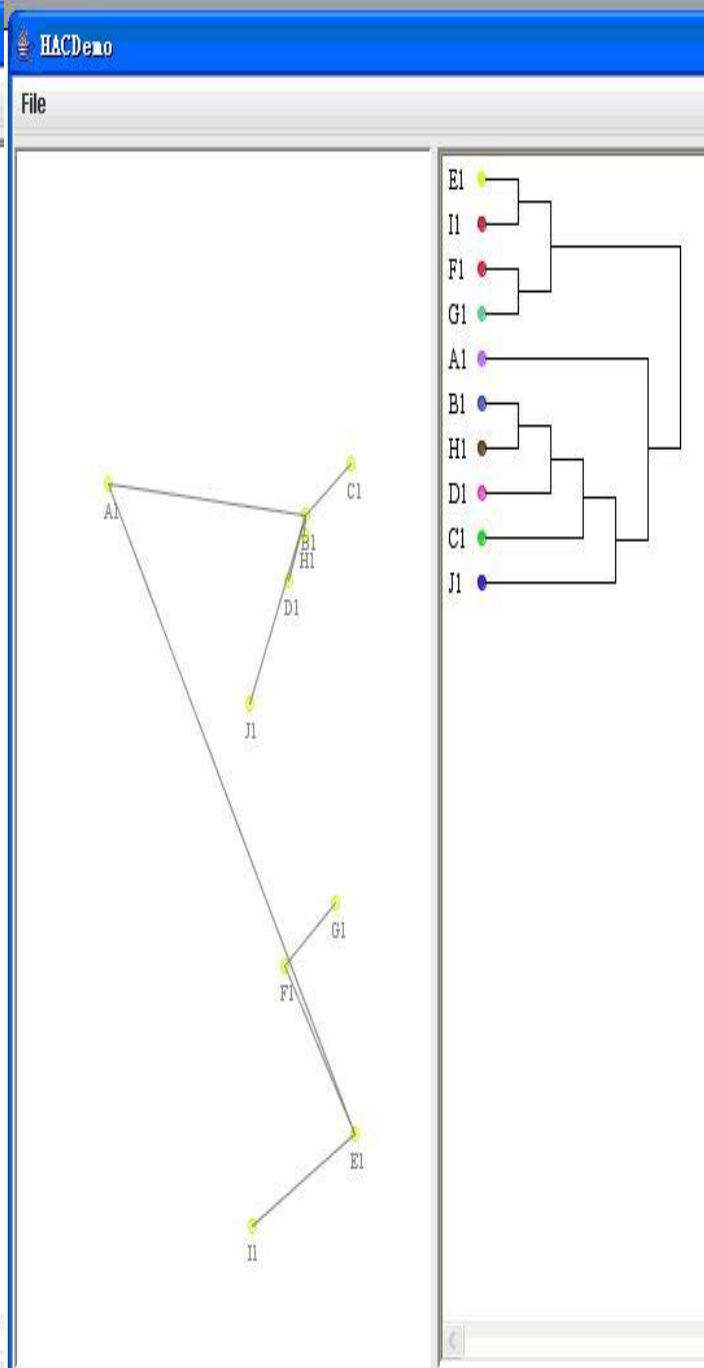
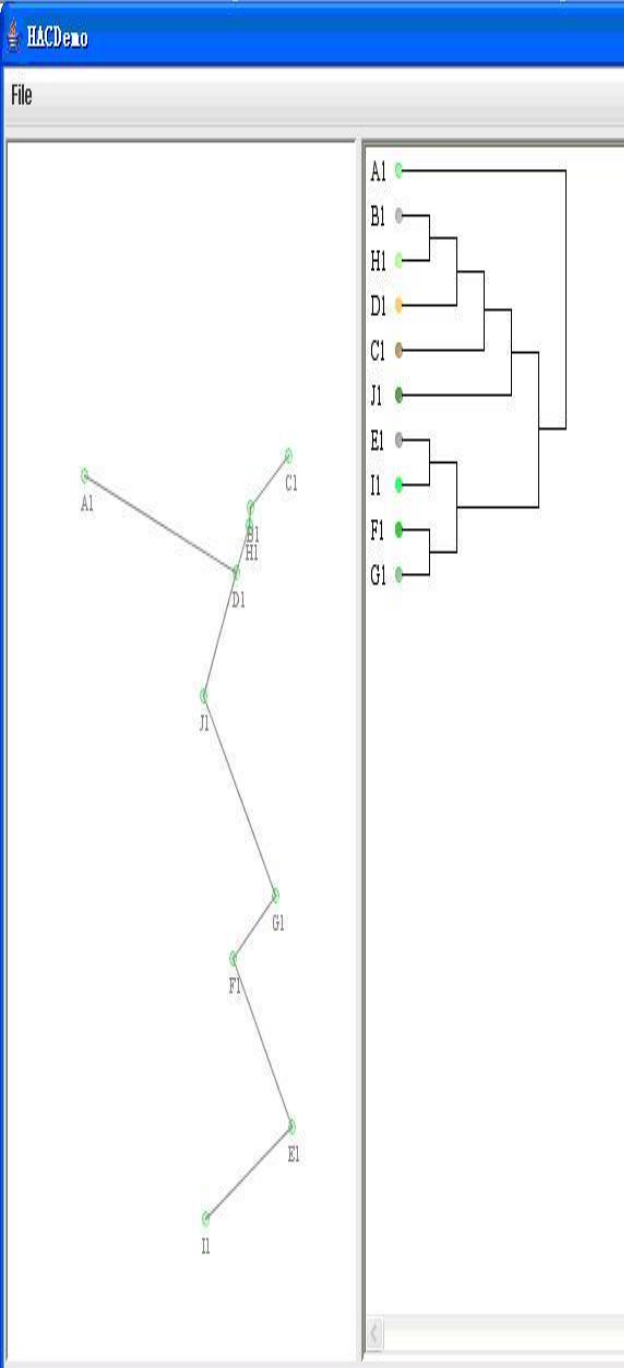
► **Figure 17.2** A single-link (left) and complete-link (right) clustering of eight documents. The ellipses correspond to successive clustering stages. Left: The single-link similarity of the two upper two-point clusters is the similarity of d_2 and d_3 (solid line), which is greater than the single-link similarity of the two left two-point clusters (dashed line). Right: The complete-link similarity of the two upper two-point clusters is the similarity of d_1 and d_4 (dashed line), which is smaller than the complete-link similarity of the two left two-point clusters (solid line).

3.5 Thesaurus及term自动关联



► Figure 17.4 Outliers in complete-link clustering. The five documents have the x-coordinates $1 + 2\epsilon$, 4 , $5 + 2\epsilon$, 6 and $7 - \epsilon$. Complete-link clustering creates the two clusters shown as ellipses. The most intuitive two-cluster clustering is $\{\{d_1\}, \{d_2, d_3, d_4, d_5\}\}$, but in complete-link clustering, the outlier d_1 splits $\{d_2, d_3, d_4, d_5\}$ as shown.





3.5 Thesaurus及term自动关联



Comments on hierarchical clustering

- * No need to specify the number of clusters in advance.
- * Hierarchical nature maps nicely onto human intuition for some domains
- * They do not scale well: time complexity of at least $O(\sqrt{n})$, where n is the number of total objects.
- * Like any heuristic search algorithms, local optima are a problem.
- * Interpretation of results is (very) subjective.