基于聚类的网络舆情热点发现及分析*

王 伟 许 鑫
(华东师范大学信息学系 上海 200241)

【摘要】根据对网络舆情分析的需求,构建出基于聚类的网络舆情热点发现及分析系统。通过对样本网页文本的特征提取,构建向量空间模型,使用 OPTICS 算法获取网页热点簇,根据热点簇特征向量对网页进行二次聚类,从而获取关于舆情的时间演变模式,为相关领域研究提供决策支持。通过二次聚类,提高舆情网页相关度的质量,使网络舆情分析更为准确可靠。

【关键词】网络舆情 热点发现 舆情分析 文本聚类
【分类号】G353.1

Online Public Opinion Hotspot Detection and Analysis Based on Document Clustering

Wang Wei Xu Xin
(Department of Informatics, East China Normal University, Shanghai 200241, China)

【Abstract】According to the requirement of online public opinion analysis, this paper builds an online public opinion hotspot detection and analysis system based on document clustering. It builds vector space model by abstracting document features from sample Web pages, and get the hot - spot cluster by OPTICS algorithm. According the vector of hot - spot cluster, the Web pages are clustered for the second time. At last, it gets the time evolution mode about the public opinion to afford decision support for specific field, and improves the quality of page correlation and analyze the public opinion more accurately.

【Keywords】Online public opinion Hotspot detection Public opinion analysis Document clustering

1 引 言

随着互联网的日益普及,中国互联网络信息中心(CNNIC)2008 年 6 月发布《第 22 次中国互联网络发展状况统计报告》数据显示:截至 2008 年 6 月底,中国网民数量达到 2.53 亿,网民规模跃居世界第一[1]。网络越来越成为人们获取与发布信息的主要渠道,网络舆情信息的导向作用愈来愈大。网络信息庞杂多样,虽然对社会的发展起了积极作用,但同时也产生了随之而来的信息内容安全问题,反动、淫秽、迷信等有害信息在网络中的传播,严重危害了国家的安全和社会的稳定。另一方面,十六届四中全会做出的《中共中央关于加强党的执政能力建设的决定》中提出,"建立舆情汇集和分析机制,畅通社情民意反映渠道",反映了党对舆情研究重要性的认识。如何在网络舆情信息采集的基础上进行舆情汇集,发现热点,并对关注热点加以跟踪分析,保障信息安全,引起了广泛关注。

收稿日期:2009 - 01 - 12
收修改稿日期:2009 - 02 - 02
* 本文系教育部人文社会科学研究项目"互联网舆情信息分析与管理机制研究"(项目编号:08JC870003)的研究成果之一。

2 研究现状

国外热点发现与分析研究较为有名的如美国的 TDT(Topic Detection and Tracking)研究项目,用以应对日益严重的互联网信息爆炸问题,对新闻媒体信息流进行新话题的自动识别和已知话题的持续跟踪。国内较为出色的系统有北大方正技术研究院的智思舆情预警辅助决策支持系统,它成功地实现了针对互联网海量舆情自动实时的监测分析。其他国内产品包括 Autonomy 网络舆情聚成系统、TRS 互联网舆情信息监控系统等。网络舆情在实践上的研究主要集中在中文信息处理与数据挖掘领域,这两个领域从不同角度对网络舆情进行研究,同时又相互交叉、渗透、借鉴。在中文信息处理领域,涉及到的内容有未登录词的识别、中文分词技术、多维向量空间对文章主题的测度等多个方面,传统上常采用词频统计的方式研究。在数据挖掘领域,涉及到的内容有舆情信息采集、自动分类、自动聚类、智能检索等方面,常采用网页特征统计的方式研究。仅聚类分析又可分为概念语义空间与相似性度、基于支持向量机与无监督聚类相结合等方法[2],在聚类方法上又可分为 K - means、BIRCH、OPTICS 等算法。在对网页主题的获取方面,即可对标题进行单独分析,又可对部分或全部全文进行分析。在以上的内容中,如何将各领域各方法有机结合,在网络舆情分析实践中取得较好效果和较高效率,是本文需要考虑的问题。

传统的网络舆情网页的统计通常基于手工主观臆断或基于网页的词频统计,手工统计效率较低,对于海量数据则难以完成,基于网页词频统计准确度较低,对网页主题没有清晰的把握,有可能造成相关网页统计数据的失真。本系统通过对相关舆情网页的二次聚类,从而在把握网络舆情主题的准确度和效率上都有较大优势。

对文章标题或关键词进行一次聚类是较为传统的舆情分析方法,取得了较好的效果。但在网络舆情实践过程中,以上方法存在一定局限性,首先网页关键词的获取相对困难,其研究工作在进一步发展中。其次网页标题有时并不能完全概括文章的主题,在对网络舆情热点特征的把握上与全文聚类相比相对较差。如果对网络舆情全文聚类又产生计算量大幅度上升的弊端,因此本文采取二次聚类的方法,即通过小范围网页一次聚类获取网页热点簇,提取舆情热点特征词,二次聚类对网页进行跟踪分析。同时,将目标的实现分为两个阶段,第一阶段为网络舆情热点的发现,第二阶段为网络舆情热点的跟踪分析。两阶段的分析方法一方面可以对舆情热点有更详细的中间数据资料,另一方面更有利于为下一阶段舆情预警工作打好基础。

3 方案设计

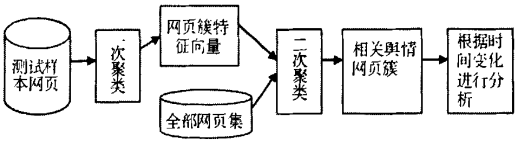

图 1 系统流程图

如图 1 所示,本系统对相关舆情网页采集后,随机抽取网页样本集合进行聚类,得到多个热点网页簇,选取关注的单个网页簇进行特征词抽取后,对全部网页使用抽取后的网页特征向量进行二次聚类,得到相关度较为纯粹的网络舆情网页集,从而描绘出网络舆情演变趋势变化。

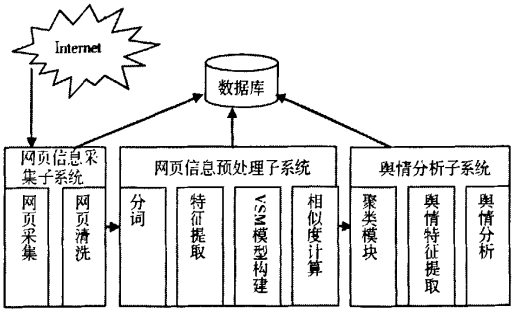

图 2 系统总体框架

图 2 展示了系统总体框架结构,本文构建出基于二次 OPTICS 聚类的网络舆情热点发现及分析系统。本系统的整个工作流程为:网络爬虫模块通过 Deep Web 技术动态查找所需新闻网页,通过网页清洗模块对网页的导航、广告、版权说明等噪声信息进行去噪后,使用节点智能识别技术抽取出相关文档字段,从而完成数据清理。分词模块对标准化文档进行特征词切分。首先随机抽取部分网页作为测试样本以获取一次特征词,特征抽取模块对样本网页进行特征词评估后,抽取出测试样

本共同的特征向量,使用特征向量的 TFIDF 值来表达文本主题,从而由 VSM 构建模块建立向量空间模型。主题发现子系统对文本关系矩阵进行 OPTICS 聚类,获取网页热点簇。此时各热点簇内网页相关度较大,二次聚类模块通过对各网页热点簇进行特征抽取后,对全部网页进行二次聚类,进而获取网络舆情热点时间演变模式。采用以上方法,解决了网页数量巨大,内容杂乱,更新迅速,不易保存所造成的对网络用户及时、准确获取所需信息的困扰,为相关领域与行业的人员提供决策支持。本系统采用中间数据多次存储处理,使得各个模块间耦合度小,内聚性高,同时也便于第三方对数据的核实与抽取和历史数据的收集。

4 系统实现

4.1 网页信息采集与清洗

本系统采用自主开发的分布式协同爬虫,可动态配置爬虫服务器数量以及爬虫数量,在不同的采集需求下动态增减使用在采集上的计算资源。系统通过网页采集子系统中的爬虫模块在 Internet 上获取网页,可对爬虫模块设置爬虫的数量、抓取速度、起始 URL、符合采集要求的 URL 的正则表达式、爬虫线程终止条件等约束,来获取相关的网页信息。对获取的网页,通过网页清洗模块清除网页中的广告、导航信息、图片、版权说明等噪声数据,萃取出相关网页的标题、正文、链接地址、采集时间等数据,导入数据库。

根据本文的实验示例,针对新闻网页,本文设计出以下字段:

字段代码	Anchor	Summary	URL	Time	Entry type	Source type
字段名	标题	正文	链接地址	采集时间	采集方式	网站来源

数据库生成字段的 SQL 语句可表示为:

```
CREATE TABLE IF NOT EXISTS 'VSMModel' (
  'anchor' VARCHAR (50) NOT NULL,
  'Summary' TEXT NOT NULL,
  'URL' VARCHAR(50) ,
  'time' DATE,
  'Entrytype' VARCHAR(50),
  'sourcetype' VARCHAR(50),
PRIMARY KEY ('anchor')
) ENGINE = MyISAM;
```

4.2 网页信息预处理

网页数据的预处理子系统主要包含网页文本分词模块、特征提取模块、VSM 模型建立模块、网页相似度计算模块。

(1)网页文本分词模块。中文分词研究已较为成熟,根据是否使用切分词典,可分为有词典切分和无词典切分。根据切分的具体方法,可分为基于规则的方法和基于统计的方法。本文采用中国科学院计算技术研究所研制的汉语词法分析系统 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System),ICTCLAS 采用了层叠隐马尔可夫模型(Hierarchical Hidden Markov Model),主要功能包括中文分词、词性标注、命名实体识别、新词识别;同时支持用户词典,其分词速度单机可达 996KB/s,分词精度到达 98.45%,分词效果良好[3]。考虑到系统平台独立性,本系统采用 ICTCLAS 的 JNI 接口。

(2)特征提取模块。首先采用抽取样本网页的全部网页词库作为网页的特征向量待选集合,由于分词后的特征向量空间维度很大,因此有必要对网页特征向量进行降维处理。本文首先根据词性进行初步筛选,笔者定义助词、介词、连词等虚词以及词语长度为 1 的无实际含义词为停用词,然后构造网页主题评价函数,对每个特征向量进行评估,选择符合预定阈值的词作为网页的特征向量集。由于对网页内容先验知识的缺乏,本文采取了词频与信息增益(Information Gain,IG)相结合的方法,将特征 T_k 信息增益的定义为[4]:

$$Gain(T,T_k)=I(T)-I(T,T_k)$$

$$=-\sum_{k=1,\dots,n}P_k*\log P_k-\sum_{k=1,\dots,n}\frac{P_k}{P_k*\log P_k}*P_k*\log P_k$$

其中 n 为特征集的维数,网页特征词语频率为 freq(k),P(k)为该特征向量的出现概率:

$$P_k=\frac{freq(k)}{\sum_{k=1,\dots,n}freq(k)}$$

特征 T_k 的信息增益值越大,说明特征 T_k 中包含的鉴别信息就越多,选择信息增益值的前 15% 作为网页特征向量,对于符合阈值要求的特征词作为网页的主题特征。

(3)VSM 模型建立模块。对于每个网页,设 T 为其特征集,$k_{i,j}$ 为网页 i 中的第 j 个特征词语,$w_{i,j}$ 为网页 i 第 j 个特征权值,m 为网页 T 中特征词语的总个数,则网页 T 可表示为:

$$\vec{Ti}=[(k_{i,1},w_{i,1}),(k_{i,2},w_{i,2}),\dots,(k_{i,j},w_{i,j}),\dots,(k_{i,m},w_{i,m})]^T$$

使用 $TF_{i,j}*IDF_i$(Term Frequency - inverse Docu-

ment Frequency)值来表示 $w_{i,j}$。

其中 TF 指 Term Frequency,表示词条 j 在网页 T_i 中出现的次数,用 Frequency(Term)表示,IDF 表示反比文档频率,用 N 表示网页集合中所有网页的数目,n_i 表示整个网页集合中出现词条 i 的网页的数目,则 $w_{i,j}$ 可表示为:

$$w_{i,j} = TF(i,j) * IDF_i = frequency(term) * \ln(\frac{N}{n_i})$$

由此,系统构建 VSM 向量模型,具体操作中,对于每个网页,采用一个散列映射表变量与之对应,由此形成词、权重值的对应关系,其网页 i 变量定义为:

Hashmap < String, Double > page[i] = new Hashmap < String, Double > ;

(4)网页相似度计算模块。对于网页相似度计算,采取了比较传统的夹角余玄值度量网页 d^i,d^j 的相似度:

$$S_T(d^i,d^j) = \cos(T^i,T^j) = \frac{\sum_{k=1}^{n} w_k^i * w_k^j}{\sqrt{\sum_{k=1}^{n}(w_k^i)^2}\sqrt{\sum_{k=1}^{n}(w_k^j)^2}}$$

4.3 网络舆情分析

该子系统包括 OPTICS 聚类模块、舆情特征提取模块、舆情分析模块。

鉴于网络舆情数据信息的离散性特征造成标志网络主题的特征向量混杂在一起,大量不能代表最后所形成各自网页簇主题特征的特征词由于聚类前各网页簇的混合而不能区分;另一方面,特征向量的混合也从客观上加大了系统的计算量,大大增加了网页计算的时间复杂度。因此针对网络舆情聚类的实际应用,本文采用多重聚类来解决以上问题。

通过选择一段时间样本网页进行一次聚类,形成多个网络舆情热点簇,通过热点簇的分析可以得到各热点簇的特征向量,此时该特征向量相对于本热点主题的相关性大大提高,使用这些特征向量对更大范围的网页数据进行二次聚类,则可以获得质量较高的舆情热点走势图表。同时,通过一次聚类和二次聚类的区分,可以将系统的工作界定为热点发现和热点跟踪。

(1)一次聚类与热点发现。对于网页信息,由于网络数据的发散性,采取密度聚类的算法,从而克服基于距离的算法只能发现"类圆形"聚类的缺点,可发现任意形状的聚类,且对噪声数据不敏感。考虑到对网页先验知识的不足,采取了对输入参数不敏感的 OPTICS(Ordering Points To Identify the Clustering Structure)算法,它通过引入核心距离(Core - distance)和可达距离(Reachability - distance),并不明确产生一个聚类,而是为自动交互的聚类分析计算出一个增强聚类顺序,OPTICS 聚类算法过程伪代码描述如下[5]:

```
OPTICS(Page, ε, MinPt, OrderList)
{ for each P ∈ Page
  let the first seed to be P;
    if P has not been processed
{ get P's neighbors and write P into OrderList
      if P is a core page
  update Ordered Seeds with each P's neighbor and check each
    reachability - distance;
  remove P from OrderedList;
    }
}
```

(2)二次聚类与热点跟踪。由于网络舆情信息的随机、复杂等特性,具体表现为网页文本大小不一,舆情主题多且杂乱,形成的舆情热点簇网页数非常不均衡。第一次聚类并不能得到关于网络舆情的精确分析,因此采用了二次聚类的方法。

首先提取出所关注网页簇中的特征词。由于经过聚类后的网页簇密度较大,各簇都有一定数量的网页来表达网页簇热点的主题,因此从网页簇中提取特征词来代表热点特征更为清晰。由于噪声网页大幅度减少,使得原来易受干扰的网页的本质特征更易被显现出来,表达网页簇主题的特征词更加纯粹,因此采用二次特征提取获取代表网页主题的特征词。

根据提取的网页簇特征词,再次对采集的全部网页进行二次聚类。由于第一次聚类所选网页范围较为广泛,某些特征词选取与所关注热点的主题无关,因此聚类质量不太理想,通过热点簇的形成,提取出与舆情热点主题相关度较大的特征词。通过使用所关注热点的特征词作为特征向量,对更大空间或时间范围的网页进行二次 OPTICS 聚类[6],重用第一次聚类模块,获得关于相关热点舆情信息,根据相关度质量较高的舆情网页,绘制出网络舆情时间演变趋势图表。

通过二次聚类,其对舆情热点的区分具有以下优势,作为舆情热点的特征向量更加集中,由于具有干扰作用的其他舆情热点特征词基本被过滤,关于舆情热点的主题凝聚性更高,特征向量维度大幅下降,处于低维空间的网页主题更易显现。

5 系统优化

5.1 分词优化

网络信息爆炸式增长同时带来了网络词汇的发展,网络新词与专有词给分词都带来困扰,尽管使用较完善的分词系统,仍有部分词汇的切分准确度较低,因此,本系统采取再次统计的方法进行分词确认[7]。通过分词后的词语集前后结合,如果结合后的新词词频超过设定值,则认为该词为新词,使用符号/new 表示。

如:奥巴马在4日美国总统选举中获胜,当选美国第56届总统。

分词后表现为:奥/b 巴马/ns 在/ p 4 日/t 美国/nsf 总统/n 选举/vn 中/f 获胜/vi ,/wd 当选/v 美国/nsf 第56/m 届/q 总统/n 。/wj

经统计,网页集中"奥巴马"词频达到了3 654次,则可视情况将该词合并为新词。合并后表现为:奥巴马/new 在/ p 4 日/t 美国/nsf 总统/n 选举/vn 中/f 获胜/vi ,/wd 当选/v 美国/nsf 第56/m 届/q 总统/n 。/wj从而避免在之后的操作中产生歧义。

5.2 近义词优化

对于分词后的词汇集,其近义、词汇缩写问题较大干扰了热点发现的准确性,降低了热点发现系统的性能。为了避免这种情况的产生,本文采取了对特征词间的语义相似度计算。在具体操作中,采用"基于知网的词汇语义相似度计算"软件包[8],对于特征词间大于k(0 = < k < = 1)的词汇组,认定为近义词,进行词语合并,并赋予一定权值 n(n < = 1)替代原词。

如:sim(很,非常) = 0.861 111,即可用新造词 < 很,非常 > 代替原词,并对原词的 TFIDF 值赋予一定的权值 k'。

对于缩写词,经计算相似度为1,则可直接用原词代替。

如:sim < 奥运会,奥林匹克运动会 > = 1,则直接用新造词 < 奥运会,奥林匹克运动会 > 代替。

6 实验示例

CNNIC 调查数据显示,网民经常使用的网络服务66.3% 以"浏览新闻"为主。网络新闻使用率为81.5%,用户规模达到2.06亿人,在网络应用中排名第二,仅次于网络音乐使用率。因此本文选取网络新闻作为系统实验示例。共采集新华网10月1日至11月30日新闻共37 805条。

首先随机选取11月1日至10日全部新闻共5 961条,作为网页簇特征词获取的测试样本。因为网页来自于现实网站,网页数据具有一定的复杂性和随机性。经过分词处理后,共计44 102个词汇,对其进行停用词处理后,得到34 565个词汇进行信息增益计算,取前15%词汇即5 185个作为网页文本的特征向量,其信息增益值较大的词汇词频如表1所示:

表1 特征向量词频

记者	美国	经济	工作	发展	公司	11月	国家	问题	市场
7 296	6 457	6 337	5 697	5 240	4 237	4 009	3 964	3 950	3 883
企业	国际	社会	金融	奥巴马	政府	总统	建设	发展	北京
3 876	3 783	3 734	3 686	3 654	3 589	3 470	3 245	2 758	2 724

通过 TFIDF 值计算,完成 VSM 模型建立,例如随机抽取的新闻网页"金融危机下中国创业者应充分利用中国市场优势"为例,其部分特征向量权重值如表2所示:

表2 网页特征向量权重值

中国	美国	出口	贸易	房地产	优势	商业	大赛	经济	欧洲
0.0774	0.0747	0.0446	0.043	0.0392	0.0388	0.0339	0.0337	0.0335	0.0332
578985	062438	756722	44482	207961	744871	295653	587 895	959987	792393
大学	国内	市场	制造业	需求	衰退	领域	消费	吸引	降低
0.0332	0.0304	0.0270	0.027	0.0264	0.0259	0.0229	0.0218	0.0210	0.0208
571889	715079	274219	00653	808337	353158	509474	926218	813371	804435

通过相似度计算与 OPTICS 聚类,获取部分热点集合。表3展示了通过聚类获取的排序前三位的新闻热点簇和与部分热点簇相关的特征词表。

表3 新闻热点主题

	网页数	特征词示例
金融危机网页簇	1 037篇	经济,国际,金融,危机,资金,股市,出口,股票,华尔街,次贷,下降,全球,…
美国总统大选网页簇	430篇	美国,奥巴马,总统,选举,麦凯恩,投票,候选人,大选,选民,共和党,民主党,竞选,当选,…
陈云林访台网页簇	237篇	两岸,关系,陈云林,海峡,< 海峡两岸关系协会,海协会 >,台湾,江丙坤,海基会,会长,…

随机选取"美国总统大选"这一热点事件作为演示网络热点跟踪分析的示例,采用网页簇生成的特征词对10月1日至11月30日所有新闻网页进行二次聚类。经统计,共获取686篇相关新闻网页,时间跨度为10月1日至11月20日,由于4日之后总统大选事件进入尾声,经过统计,11月20日之后相关网页数基本为零。如图3所示,在11月4日左右,由于4日美国第56任总统的揭晓,关于"美国总统大选"的新闻报道

达到了顶峰,与网络舆情的时间演变模式相符。根据时间演变模式,可以有针对性的进行趋势分析与预警系统的构建,为各行业人员提供决策支持。

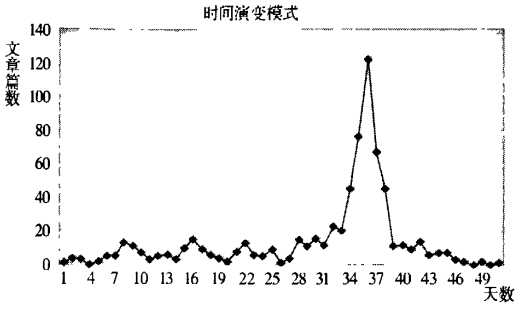

图 3 时间演变模式

7 结 语

本系统基本达到了文章提出的目标,在网络舆情分析的精度与效率方面有较大提高,然而系统中各算法、参数和阈值的恰当选择是个长期的过程,同时,本系统的时间复杂度与算法优化方面还有较大的发展空间,只有在实践中不断调整,在不断与系统用户反馈交流中才能达到最佳。

同时本系统在以下方面值得进一步研究。

(1)在系统指标与参数的选取方面。关于网络舆情的分析还可参考更多指标进行测评,如网页发布者的影响力、网站的权威度等,从而更真实的把握网络舆情的发展状况。同时,本系统只作了词语的近义优化而未考虑词语多义问题,人称指代问题也需作一些改进。

(2)定性分析方面。对网络舆情的定性分析也是非常必要的,针对不同的舆情热点结果,需采取不同的策略来解释图表所展示的结果。通过对网络舆情态势的判别,分析出该舆情发展变化的类型,从而把握其发展脉络与趋势。

(3)系统应用范围方面。本系统通过聚类获取了关于某一具体舆情事件的特征向量,这些特征向量描述了该舆情事件的主要轮廓。因此,这种特征向量对事件描述的方法同样可以应用到对热点人物的描述,同时可能会产生其他一系列聚类细节问题。

(4)舆情热度的评判标准方面。鉴于某些网页的报道与热点相关性并非唯一,可能出现一个网页与多个热点相关,而以文章篇数作为舆情热度的评判标准可能存在部分失真。本系统根据文章相关度进行网页聚类,则可考虑采用网页与特征向量的相关度值来作为舆情分析标准,采用与热点的相关度模糊标准来表示时间变化趋势,将会展现出更加精确的网络舆情变化趋势。

参考文献:

[1] 中国互联网络信息中心. 第 22 次中国互联网络发展状况统计报告[EB/OL]. [2008 - 07 - 23]. http://www.cnnic.cn/uploadfiles/pdf/2008/7/23/170516.pdf.

[2] 李晓黎. WEB 信息检索与分类中的数据采掘研究[D]. 北京:中国科学院计算技术研究所,2001:61 - 90.

[3] ICTCLAS 简介[EB/OL]. [2008 - 12 - 01]. http://ictclas.org/sub_1_1.html.

[4] 姚清耘. 基于向量空间模型的中文文本聚类方法的研究[D]. 上海:上海交通大学,2008.

[5] 孙学刚,陈群秀,马亮. 基于主题的 Web 文档聚类研究[J]. 中文信息学报,2003(3):12 - 16.

[6] 郭建永,蔡永,甄艳霞. 基于文本聚类技术的主题发现[J]. 计算机工程与设计,2008(6):1426 - 1428.

[7] 徐文海,温有奎. 一种基于 TFIDF 方法的中文关键词抽取算法[J]. 信息系统,2008(2):298 - 301.

[8] 刘群,李素建. 基于《知网》的词汇语义相似度计算[A]. 第三届汉语词汇语义学研讨会,2002.

(作者 E - mail:asdwangwei@yahoo.com.cn)