基于多语境的相关词自动提取

章成志，苏兰芳

（南京大学信息管理系，南京 210093）

摘　要： 通常词语在一定的语境下出现会更加确切表达其意思。不同的语境从不同侧面反映了词汇关系。本文利用语料库、释义词典、用户搜索日志作为识别相关词的语境，设计并实现了相关词自动提取系统。分析实验结果，我们发现，虽然面向相同的基本词汇集合，但是基于不同知识源提取的相关词之间的重复率很低，各个结果间的互补性很强，因此，结果整合非常有必要。在本系统中，通过直接整合途径得到了最后的相关词词表。

关键词： 相关词；多语境；语料；释义词典；用户日志

Automatic Extraction Relevance Terms in Multi-Context Environment

Zhang Chengzhi，Su Lanfang

(Department of Information Management, Nanjing University, Nanjing 210093)

Abstract: Generally, the meaning of a word can be expressed in the context. Different context can reflect the relationship of terms from different side. The corpus is the formal using of words, and dictionaries define the concept of words, and the User-logs involve users' indirect participation. The authors choose them as context to extract the relevance terms. From the experiment results, they find that the overlap ratio of results in different contexts is very low. So, it is necessary to integrate the different results. All of the relevance terms were integrated to a relevance table through direct integration.

Keywords: Relevance Term; Multi-Context; Corpus; Definitions Dictionary; Query Log

1　引言

在信息检索中，由于用户使用的自然语言通常不能考虑到所使用词汇的相关词，很容易造成信息漏检或误检，降低了检索效果，解决这个问题的途径之一便是查询扩展或相关词提示。应用查询式扩展查询扩展或相关词提示可以辅助用户正确表述信息需求，降低信息用户智力负担。此外，在用户查询式的基础上提供相关词，通过检索式的重新构建可以进一步完善检索式，达到扩检和缩减的效果。

查询扩展的方法基本可以分为两大类：局部分析法和全局分析法。最近几届 TREC 会议的研究结果表明，使用局部分析法的查询扩展方法，通常可以比较显著的提高信息检索效果。但也有研究表明，这些查询扩展方法的效果并不稳定，其效果强烈依赖于第一次检索的结果。通常先考虑进行全局分析法的查询扩展，在获得相对更可靠的检索效果之后，再进行局部分析法的查询扩展[1]。在局部分析法的查询扩展中，获得高可靠性的辅助资源是关键，从过去的研究来看，这种资源一般是对词汇具有一定程度的控制功能的词表或词典，依据词表构建自动化程度的不同，可将其分为手工构建词表和自动构建词表。

作者简介：章成志（1977-），男，博士研究生，研究方向为智能信息检索，zcz51@citiz.net。

苏兰芳（1980-），女，硕士研究生，主要研究方向为信息检索与信息系统。

手工构建的词表有两种，一种是通用词表，如 WordNet，这类词表并不能显著提高检索效率[2]。另一种是专门应用于检索系统的手工构建的词表。自动构建词表是应用自然语言处理技术，从文档集合或用户查询日志中中挖掘出具有一定关系的词汇，并在词汇的基础上组织词表。相关词自动识别属于词表自动构建的范畴，国内外与相关词识别的工作主要有[3]~[10]，大多都是基于单一语境来计算词语的相关度。单一语境能够提供的词汇以及词汇间的相关特征毕竟有限，因此很难构建高质量的相关词词表。由于资源电子化程度的不断提高，以及社会知识共享程度的改善，许多以前很难获取的资源现在已经可以通过各种途径共享。在这样的研究背景下，本文提出了基于多语境获取相关词的思路，即：不同的语境从不同侧面反映了词汇关系，利用语料库、释义词典、用户搜索日志作为识别相关词的语境，设计并实现了相关词自动提取系统，希望能够整合不同语境下的词汇关系，提高相关词识别效率。

2 相关词自动抽取系统的总体结构

本实验系统的设计思路是：收集语料库、释义词典和用户日志三种语境下的数据资源，将其规范化处理后分别导入数据库，经过分词、去停用词等预处理之后，应用不同的同义词识别算法从中获取相关词候选集，最后整合的相关词候选集成相关词词表。整个系统共分为多语境数据获取、多语境相关词提取以及相关词表整合三个模块。系统总体流程如图 1 所示。

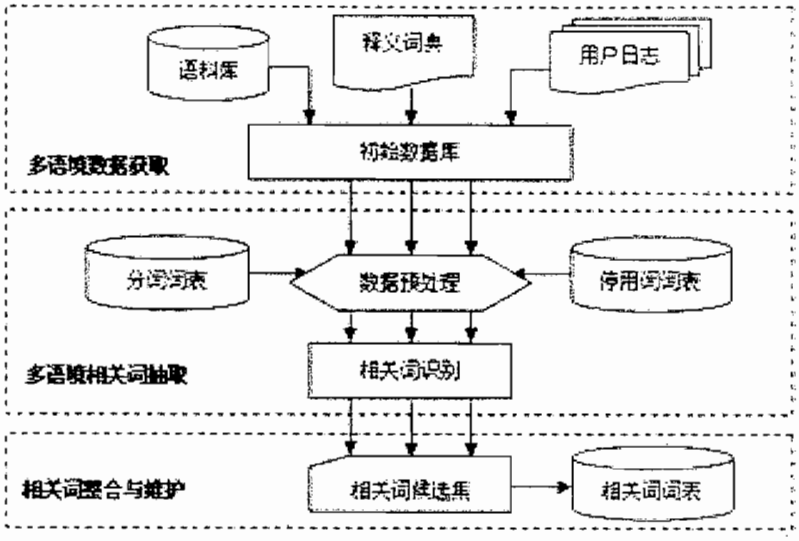

图 1 系统流程图

2.1 多语境数据获取

我们收集到的资源没有规范化的格式，为了方便以后的数据处理，我们首先要将其导入到规范化的数据库中。在数据导入过程中，首先分析源数据的存储格式，提取源数据内容存入相应数据字段。

2.2 多语境相关词提取

语料库和释义词典都是以句子为基本单位，要想从中提取相关词，首先要分词。虽然目前为止，中文分词研究已经取得了一定的进步，但是因为汉语言本身的复杂性，分词准确度依然有待提高。在我们的实验系统中，以《人民日报分词词表》[11]为分词词表，采用逆向最大匹配分词算法，将句子分解为基本词汇。考虑到有些词汇对于相关词提取没有意义，为了提高系统效率，我们在分词之后进行了停用词处理，去掉了没有实际意义的停用词。预处理完毕后再进行相关词提取，包括从大规模语料库中提取相关词、从释义词典中提取相关词和从用户日志中提取相关词。三个过程互相独立，分别实现相关词提取功能。

2.3 相关词整合与维护

如前所述，不同的知识源从不同侧面反映了词汇间关系，那么基于不同知识源提取的相关词结果必然会有所不同。因此，在获取了基于各种知识源提取的相关词结果后，我们对结果进行了整合，最终得到我们提取到的相关词词表。

3 相关词自动提取方法实现描述

3.1 语料库相关词提取方法描述

实验系统中用到的语料库数据《经济日报》从 1983 到 2003 年的全文数据，共 112754 篇，约 1 亿汉字。从语料库中挖掘相关词，是基于词频共现原理，共现窗口选定为文档中的自然段落。从大规模语料库中提取相关词，可以转化为依据统计值判断词汇间的紧密程度，如果统计值表明两个词汇间具有很高的紧密度，那么就可以认为它们是相关词。

在基于字串内部结合紧密度的汉语自动抽词实验中，罗盛芬和孙茂松[12]比较系统的考察了九种常用统计量在自动抽词中的作用。实验过程中，对所有统计量，都采用相同的测试集和抽词过程，计算结果表明，互信息的效果最好。在本实验系统中，我们借用互信息来度量词汇间的相关关系。

在我们的试验中，选用分词词表中的词作为特征词 w_i，则特征词集合 A 可表示为：$A=\{w_1,w_2,\ldots,w_i,\ldots w_n\}$。由语料库中文档的自然段落 T_j 组成的段落集合 T 表示为：$T=\{t_1,t_2,\ldots,t_j,\ldots,t_m\}$。将互信息公式应用于相关词挖掘，对于词汇 w_i 和 w_j 之间的互信息计算公式如下：

$$MI(w_i,w_j)=\log\frac{P(w_iw_j)}{P(w_i)\times P(w_j)} \qquad (1)$$

其中，$P(w_i)$、$P(w_j)$分别为特征词 w_i 和 w_j 在段落空间 T 中出现的概率，$P(w_i,w_j)$是二者在空间 T 中的同现概率。根据最大似然原理，公式(1)转化为如下形式：

$$MI(w_i,w_j)=\log\frac{f(w_iw_j)/N}{(f(w_i)/N)\times(f(w_j)/N)}=\log\frac{N\times f(w_iw_j)}{f(w_i)\times f(w_j)} \qquad (2)$$

其中，f（w_i）和 f（w_j）分别是特征词 w_i 和 w_j 在段落集合 T 中出现的总次数，$f(w_i,w_j)$是二者在 T 中的共现频次，N 是集合 T 中的元素总数，也就是语料库的规模。一般来说，两个特征词在语料库中共同出现的频率越高，二者相关性越高，但是，互信息不仅仅与共现频次有关系，还受语料库大小的影响。词汇间互信息的大小反映了词汇间关联紧密程度，互信息值越大，关联的越紧密，是相关词的可能性越大。我们设定一个阈值 R，若 MI（w_i, w_j））≥ R，则 w_i 和 wj 进入相关词候选集。通过调整阈值 R，可以限制相关词对的个数。

3.2 释义词典相关词挖掘算法描述

释义词典数据来自于《现代金融词典》[13]，共1001条词汇。经过数据净化和数据标准化后，形成规范的文本格式，每一行表示一个词条。本实验系统基于释义词典的相关词挖掘算法中，只基于第一个假设，为每个词对构建词汇向量空间模型。在构建的向量空间模型中，词汇释义中的每个词作为一个维，用来描述词汇空间。例如“直接金融”和“间接金融”的释义分词（经停用词过滤）结果如下：

直接金融：没有/金融/中介/机构/介入/资金/融通/方式/
间接金融：通过/金融/中介/机构/进行/资金/融通/方式/
那么它们构成的词汇空间为：{没有 金融 中介 机构 介入 资金 融通 方式}
“直接金融”和“间接金融”的向量描述分别为：
直接金融：{1 1 1 1 1 1 1 1}，间接金融：{0 1 1 1 0 1 1 1}

接下来便是判断两个向量之间的相关性，判断相关性的方式很多，我们这里选则了计算词汇间的 Cos 值。在向量空间中，Cos 值越大，说明两个向量之间的夹角越小，也就是二者相关的可能性越大。因此，给定阈值 Q，若 Cos（w_i, w_j））≥Q，则 w_i 和 wj 进入相关词候选集。同样，通过调整阈值 Q，可以限制相关词的个数。

3.3 用户日志相关词挖掘算法描述

本文选用了由国外搜索 alltheweb 网站获取的 2002 年 5 月 27 号全天的用户检索日志作为实验对象。在我们获取的用户检索日志中，共 2,277,452 条检索记录，检索词大部分为英文。每条检索记录包含四个字段：时间：包括年、月、日；IP 地址：发出检索请求的客户端主机的 IP；语种：检索词所属语种；查询式：用户的检索用词；示例如下：

20020528	i=195.246.158.67	l=english	nik putnam
20020528	i=195.246.158.67	l=english	wicca history

20020528	i=195.246.158.67	l=english	badger
20020528	i=195.246.158.67	l=english	dinkler
20020528	i=195.246.158.67	l=english	"christine albee"

我们抽取用户日志数据中的 IP 地址和查询式两个字段，将同一 IP 地址的查询式聚集成一条记录，之后用用分隔符“/”分隔查询式中的检索词，存入数据库中，处理之后共有 2,277,344 条记录。例如对于前面提到的几条日志纪录，经处理后形成记录样例如下：

195.246.158.67 nik/putnam/ wicca history/badger/dinkler/"christine albee"/

一般同一用户提交的检索词之间更容易具有相关关系，也就是说，如果两个检索词经常共现于一个查询式集合中，那么我们可以认为它们是相关的。可以把查询式集合看作段落，将所有用户的查询式作为语料库，借用词共现原理，计算词汇间的互信息，从中抽取相关词对。因为暂时未能获取比较正规的中文检索系统的用户日志，而从英文日志中获取的相关词不能参与最后的相关词整合，所以，我们以来自于《当代金融词典》中的 1001 条词汇为检索词，自动提取了百度[1]和北大天网两大大中文搜索引擎所提供的相关词。

4 实验结果分析

在本实验系统中，我们以《当代金融词典》中的 1001 个词汇为基本词汇集合，分别在大规模语料库、词汇释义和用户检索日志中判别这些词汇间的相似度，从中提取相关词，下面将分别描述和分析基于上述三种知识源的相关词提取结果。

4.1 基于大规模语料库相关词识别实验结果与分析

在基于大规模语料库的相关词提取实验中，计算出的词汇间互信息值分布于 0-11 之间，为了统计方便，我们将所有的互信息值缩小 10 倍，使其分布于 0-1.1 之间，如图 2 所示。根据实验数据的互信息值分布规律，以及对词对间关系的考察，我们最终选定阈值为 0.75，即认为互信息值大于 0.75 的词对为相关词，共得到 597 对相关词。

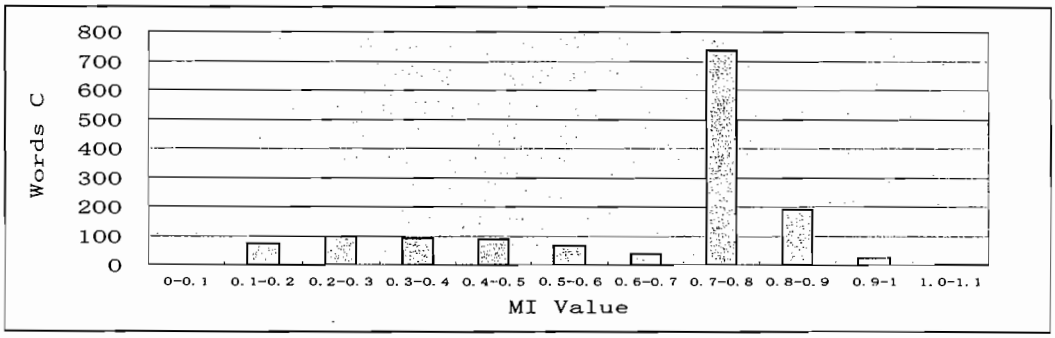

图 2 互信息值分布图

4.2 基于释义词典相关词识别实验结果与分析

本实验系统中用 Cos 值测量词对间相似度，通过调整阈值，可以控制相关词对个数。

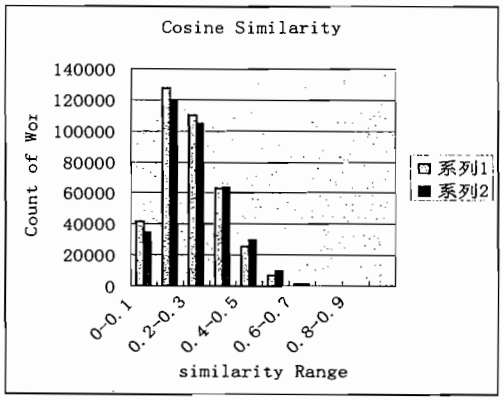

图 3 Cos 值分布图

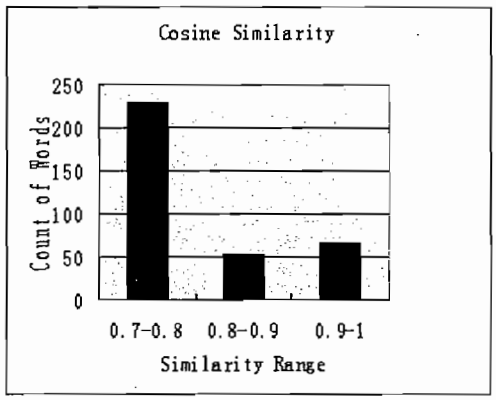

图 4 Cos 值在 0.7-1 间分布图

我们计算出《现代金融词典》中的 1001 个词汇间 Cos 值的分布如图 3 所示。图 3 中，系列 1 是包含 10 万词

条的分词词典，系列 2 是包含 14 万词条的分词词典，由此可以看出，不同的分词词典会影响相关词识别结果，但 Cos 值的总体分布趋势并没有改变。将相关词对按照 Cos 值降序排列，可以发现，随着 Cos 值的降低，词汇间相关性越来越松散。Cos 值为零，说明二者完全不相关，Cos 值为 1 的词汇间的关联度最强。在我们的实验系统中，采用《人民日报分词词表》的分词结果，因为 Cos 值小于 0.7 的词对间相关性下降很多，所以我们设定阈值为 0.7，共获取了 348 对相关词，其 Cos 值分布如图 4 所示。

4.3 基于用户日志的相关词识别结果分析

百度和北大天网两大搜索引擎是中文信息搜索的常用工具，这些网站分析大量用户的检索日志，依据特定的规则和算法，提取出与当前检索式比较相关的检索词返回给用户，辅助用户重构检索式。因此我们把它们作为基于用户日志相关词提取的实例。

实验过程中，我们以《现代金融词典》中的 1001 个词汇为基本词汇集合，以这些词汇为检索词提交给百度和北大天网，如果某个检索词返回的相关搜索词也属于基本词汇集合，则将这对相关词提取出来。分别统计了百度和北大天网返回的相关词总对数，以及在基本词汇集合中提取到相关词的词汇的总数，统计结果如表 1 所示。

表 1　百度和北大天网相关词识别统计结果比较表

项目 \ 搜索引擎	百度	北大天网
相关词对总数	**2702**	202
有相关词的词汇总数	560	198

由表 1 的统计结果可以看出，百度的相关搜索功能比较成熟，相关词覆盖率远远高于北大天网。整合百度和北大天网中提取出相关词，去重后共得到 2795 对相关词。分析整合后的相关词，具有如下特点：相关词覆盖率非常高，为大部分词汇提供了相关词；大部分都是词形相关，概念相关词对比较少；常用词的相关词范围广泛，返回的相关词对较多，而非常用词的相关词较少甚至没有。

4.4 相关词识别结果比较与整合

上文对基于各种知识源中相关词挖掘结果进行了详细的分析，它们在相关词提取方面，特点分别如下：①搜索引擎覆盖率很高，相关性也较强，但是大多是基于词形相关，而不能识别出概念相关的相关词；②释义词典中发现的相关词之间的相关性最强，但是覆盖率较低，而且会遗漏掉释义差别较大但实际上相关度很强的词对；③基于大规模语料库能够挖掘出一些从词形和释义来看都不相关的相关词，例如“支付手段”和“银行卡”。但是因为一些非常用词出新频率为零，导致很多相关词对不能提取出来。

我们对基于三种知识源的相关词识别结果进行了统计分析，结果如表 2 所示。

表 2　三种知识源相关词识别统计结果比较表

项目 \ 语境类别	大规模语料库（C）	释义词典（D）	用户日志（L）
相关词对总数(a)	597	348	**2795**
有相关词的词汇总数(b)	495	183	**616**
a/b	**49%**	**18%**	**62%**

在我们的实验中，互信息值和 Cos 值都和词对间的相关性成正比，但是比较同一词对的互信息值和 Cos 值，我们发现，有些词对的互信息值和 Cos 值大小互相矛盾。例如，“现金流通”和“现金流通量”两个词在大规模语料库中计算出的互信息值为 10.75，但在释义词典中计算出二者的 Cos 值却只有 0.36；而 Cos 值为 1 的“硬货币”和“硬通”的互信息值小于 0。而有些被搜索引擎推荐为相关词的词对，在其它两种方法中却没有识别出来，例如“个人信托”和“公益信托”在搜索引擎推荐的相关词中出现，但是其互信息值小于零，Cos 值也只有 0.47。我们进一步比较了基于三种知识源的相关词重合情况，结果发现，虽然面向相同的词汇集合，但是基于不同知识源提取的相关词结果差别很大。没有任何词对同时被三种知识源推荐为相关词，有少数词对被两个知识源同时推荐为相关词，统计结果如表 3 所示。

表 3　相关词对总数统计结果

项目 \ 语境	C	D	L	C+D	D+L	D+L
相关词对总数	597	348	2795	4	13	24

由此可以看出，基于单一知识源提取的相关词都存在片面性，而各种知识源间的互补性很强，因此，整合基

于三种知识源的相关词识别结果，能够尽可能的避免基于单一知识源提取相关词造成的遗漏，提高相关词的召回率。

相关词识别结果的整合可以通过直接整合或加权整合两种途径实现：①直接整合：充分信任每种知识源，对于两个词汇，只要有被某个知识源推荐为相关词，则将其收入相关词表。②加权整合：即给每种知识源附以一定的权重，然后依据组合值对词对重新排序，设定阈值，组合值大于阈值的才提取为相关词。直接整合的优势在于方法简单，便于操作，但是缺少对识别准确度的进一步限制。例如在我们的实验中，搜索引擎推荐的相关词总对数远远超出其它两个知识源的相关词对数，由于搜索引擎推荐的相关词大部分都只是词形相关，导致最后得到的相关词表中词形相关的相关词对占绝大多数。加权整合能够通过权重调整知识源对相关词识别结果的影响，但是最后的整合结果对权重的依赖性很强，而我们很难精确的界定每种知识源的权重。在实验系统中，我们采用了直接整合，共得到 3698 对相关词，为 685 个词汇推荐了相关词，占基本词汇集合的 68%。

5 结语

不同的语境从不同侧面反映了词汇关系，语料库是大量真实文本的集合，是词汇的正规应用；释义词典中包含的是词汇的详细定义；而用户日志有了用户的间接参与，反映了用户视角的词汇关系。这三种语境中隐含的词汇关系各有侧重。本文利用它们作为识别相关词的语境，设计并实现了相关词自动提取系统。分析实验结果，我们发现，虽然面向相同的基本词汇集合，但是基于不同知识源提取的相关词之间的重复率很低，各个结果间的互补性很强，因此，结果整合非常有必要。在本系统中，通过直接整合途径得到了最后的相关词词表基于多知识源的相关词自动识别是一个非常值得探讨的研究思路，本文所作的工作只是初步的尝试，还有很多方面需要深入的探索，今后我们需要做的工作主要有：获取某个领域内的多种语境数据，研究多语境中提取的相关词的整合方法等，另外，考虑到网络本身的变化性，即经过一段时间之后，相关词提取结果的发生变化的情况也是今后需要进一步解决的问题。

参考文献：

[1] 贺宏朝等. 一种基于上下文的中文信息检索查询扩展.中文信息学报.2002,16(6):32-37.

[2] Voorhees, E. M. Query expansion using lexical semantic relations. Proceedings of the 17th annual international ACM-SIGIR conference on research and development in information retrieval. Dublin, Ireland. 1994: 61 - 69.

[3] Yufeng Jing, W. B Croft. An Association Thesaurus for Information Retrieval. Technical Report: UM-CS-1994-017. University of Massachusetts. 1994.

[4] Crouch C. A Cluster-based approach to thesaurus construction. Proceedings of the Eleventh Annual International ACMSIGIR Conference on Research & Development in Information Retrieval, Grenoble, ACM Press, 1998: 309-320.

[5] Carolyn J. Crouch, Bokyung Yang. Experiments in automatic atatistical thesaurus construction. SIGIR 92. 1992: 77-88.

[6] Hsinchun Chen, Kevin J., Lynch. Automatic construction of networks of concepts characterizing document databases. IEEE Transactions on Systems. 1992, 22(5): 885–902.

[7] Peter D. Tumey. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the 12th European Conference on Machine Learning. Freiburg, Germany. 2001: 491–502.

[8] Pierre P. Senellart, Vincent D. Blondel, Automatic discovery of similar words, chapter in: Survey of Text Mining, Springer-Verlag, 2003.

[9] Masaki Murata, Toshiyuki Kanamaru, Hitoshi Isahara. Automatic synonym acquisition based on matching of definition sentences in multiple dictionaries. CICLing 2005, LNCS 3406. 2005: 293-304.

[10] 崔航,文继荣,李敏强. 基于用户日志的查询扩展统计模型.软件学报,2003,14(9):1593-1599.

[11] http://ccl.pku.edu.cn/doubtfire/Course/Chinese%20Information%20Processing/Source_Code/Chapter_8/Lexicon_full_2000.zip, Accessed:2006,4,20.

[12] 罗盛芬,孙茂松.基于字串内部结合紧密度的汉语自动抽词实验研究.中文信息学报,2003,17(3): 9-14.

[13] 王益,白钦先.当代金融辞典.北京:中国经济出版社,2000