Web 信息检索中相关词提示技术与评测

徐小琴　章成志

（南京大学信息管理系，南京 210093）

摘　要： 为了明确用户的查询需求，很多搜索引擎和全文数据库提供了相关词提示功能。本文简要介绍了 Web 信息检索中相关词提示的获取技术，并对相关词提示效果进行实际调查分析。从关键词库中随机抽取若干关键词，在选定的搜索引擎和全文数据库上进行信息检索，获取抽样关键词的相关提示词。通过分析，给出相关词提示在改善检索效果和用户满意度方面的综合评价。

关键词： 相关词提示；查询扩展；信息聚类；查询式专指度

Evaluation of Relevant Term Suggestion in Web Information Retrieval

Xu Xiaoqin, Zhang Chengzhi

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

Abstract: Term suggestion mechanism aims to find users' the information need, and is commonly employed in search engine and full-text databases. The authors introduce the technique of term suggestion in Web information retrieval, and analysis on the effort of term suggestion. They chose about 100 key words from the keyword database and searched in four chosen search engines and full-text databases. Through the investigation and analysis, the authors evaluate the effort of term suggestion in improving the retrieval quality and users' satisfaction.

Keywords: Term Suggestion; Query Expansion; Information Clustering; Query Specialization

1　引言

随着 Internet 的飞速发展，信息资源的分布和共享已经超越时空的限制，用户面对的信息资源库越来越丰富。到 2006 年，仅 Google 就索引了 80 亿 WebPages[1]。对于大多数课题来说，搜索引擎的返回结果数都比较大，用户要查找到需要的信息非常困难。另外，由于大部分搜索引擎用户是普通网络用户，在检索策略和检索技巧上缺乏必要的知识，提交的查询请求往往比较短。通过对微软公司旗下的 MSN 中的 Encarta 在线百科全书网站连续两个月的用户查询记录进行分析，发现 49%的用户查询仅有一个单词，33%的查询由两个单词组成，用户平均使用 1.4 个单词描述他们的查询[2]。在查询词的使用方面，由于存在同义词、多义词、歧义词和词汇孤岛等问题，用户选用的词与文献集中的词不匹配。词的不匹配现象导致检索结果的准确率和召回率不高，偏离用户的信息需求。目前，大多数搜索引擎主要是通过相关词提示帮助用户优化查询式，明确用户的信息检索需求。相关词提示是搜索引擎系统为用户提供相关词，帮助用户重新构造更加有效的查询式，从而减少多余检索步骤的检索技术[3]。常见的两种相关词提示方式是，相关搜索词和聚类浏览。笔者对 53 个中英搜索引擎进行了调查，结果表明：62%的搜索引擎提供相关词提示功能，“相关搜索词”占 45%。

作者简介：徐小琴（1982-），女，本科四年级学生，xq_xu12@hotmail.com，研究方向为信息检索与信息系统。

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

2 相关词提示的作用和方法

2.1 相关词提示的作用

基于“相关反馈”（relevance feedback）的交互查询模式，其实现方式是在前一次检索返回的文件中，选取重要的特征，反馈给系统，以期找到更多相关的数据。选取的特征若是文件本身，则可称为相关文件反馈；若为相关词，则称为相关词反馈，或检索词提示(term suggestion) [4]、相关词提示。在全文检索环境中，要判断相关文件，需要对文件做相当程度的浏览，给用户造成额外的负担。相比之下，相关词提示因为牵涉到的额外信息较少，用户较易判断，是一种比较好的查询交互方式。然而，让系统自动判断出有用的相关词，比起让系统只提供文件让用户判断，是一项复杂而困难的工作。具体实现中，相关词的选择权应控制在用户手中，由用户判断选择所需的相关词。Koenemann 的研究表明，通过相关词提示帮助用户重构的查询式，效果优于系统自动重构的查询式[5]。

2.2 相关词提示的方法

相关词提示的原理是，搜索引擎通过聚类技术获取与查询式相关的词，经过相关度计算，将排在前面的相关词返回给用户。其技术背景是信息检索领域的信息聚类技术。相关词提示的形式主要有两种：一种是在检索结果页面的上方或下方提供“相关搜索”词，如百度[6]，另一种是在检索结果页面的左侧提供聚类浏览导航体系，如 Vivisimo[7]。

2.2.1 相关词提示的技术背景——信息聚类

聚类（Clustering）是指把没有分类的事物，在不知道应分成几类的情况下，根据事物彼此不同的内容属性进行辨认，将具有相似属性的事物分为一类，使得同一类的事物具有高度的相似性，最后确定每个事物所属类别[8]。相关词提示是信息聚类的一个应用。相关搜索词，通常是以用户查询日志中的历史查询式为对象，采用聚类技术计算关键词之间的相似度，返回相关度高的相关词。而聚类浏览是以系统返回文献为对象，依据其内容进行聚类，以相关词为类别创建类目，使用户能够从更高主题层次上查看搜索结果。

2.2.2 相关词提示的实现技术

（1）“相关搜索词”的获取技术

“相关搜索词”获取的基本方法是利用用户的 Web 查询日志获得查询词的相关词。即：在用户查询记录的基础上建立用户查询空间，针对每一个查询词，记录用户曾经点击查阅的文献，从用户点击查阅的文献中抽取特征关键词进行相似度计算[2]。该方法充分利用用户的历史查询空间，参考用户对每次查询结果的反馈情况。其相关依据并不依赖于初始检索的结果，也不需要用户给出及时的反馈来判断，而是建立在对大量用户长期信息搜索和浏览行为的分析和学习的基础上，得出具有普遍意义的相关度量。

（2） 聚类浏览技术

聚类浏览是把搜索引擎返回的查询结果依据其内容进行聚类，以相关词为类别创建类目体系，使同类中文献之间具有较大的相似性，而类与类之间的文档具有较小的相似性。然后把类目呈现给用户，使用户能在更高的主题层次上来察看搜索引擎返回的结果。它不需要用户的及时反馈，而是针对系统返回的所有结果进行聚类[8]。

（3）相关词提示技术比较

上述两种相关词的获取方法，前者是基于查询日志的相关词获取方法[10]，后者是基于文本集合的相关词获取方法[11]。前者是从用户角度出发，基于用户的历史查询记录，依据的是词与词之间的相似度。抽取的相关词贴近用户语言（查询式），能够获取概念相关的关键词，其局限性是用户往往只点击排列在前面的结果记录，统计过程中忽略了大量排在后面的网页，如果排在前面的网页不符合用户的信息需求，会大大降低相关词提示的质量。基于文本集合的相关词获取方法，没有考虑用户的历史查询记录，仅仅针对当前检索行为从结果文献中抽取关键词，会出现大量偏离用户信息需求的不相关词，无法表达概念相关的关键词。但它不依赖于用户的查询日志，技术比较成熟，依据的是文献之间的相似度。因此，将这两种方法结合起来，能提高相关词提示的性能[2]。

3 相关词提示评价

本文选取一些关键词作为查询样本，在选定的搜索引擎和全文数据库上进行检索，获取查询词的相关词，根据相关性评分标准进行人工评分，对统计获得的数据进行“查询扩展”和“查询式专指度”分析，给出相关词提示在改善检索效果方面的综合评价。

3.1 评测步骤

具体评测分为五个步骤：①从事先生成的关键词库中，随机选取关键词作为查询样本；②选取适当的搜索引擎和全文数据库作为调研对象；③用样本关键词进行信息检索，获取其“相关词”并进行数据统计；④设计相关性评分标准，并对每个查询词的“相关词”进行人工评分；⑤利用步骤③和④获取的统计数据，进行“查询扩展”分析和“查询式专指度”分析，并给出相关词提示在改善检索效果方面的综合评价。

3.1.1 选取调研数据源

本论文调研数据来源于 CSSCI[12]（中国社会科学引文索引）2004 年度经济类关键词（共 1623 个），按测试方法分为封闭性测试数据和开放性测试数据。封闭性测试数据，是从来源数据库中选取 20 对关键词，每一对都是经过人工确定的相关词，即有很高的相关度。目的是考察搜索引擎和全文数据库在相关词方面的覆盖率。具体步骤是，从每一对关键词中取一个词作为查询词，提交给搜索引擎/数据库，考察另一个词是否包含在搜索引擎和全文数据库提供的相关词中。开放性测试数据，是从来源数据库中自由抽取 60 个查询词，进行信息检索并获取其相关词的统计数据。目的是在不清楚查询词有没有相关词的情况下，考察搜索引擎提示的相关词，并对相关词的效果进行评价。从查询式专指度出发，把查询词分为单个词和组配词两类，其中单个查询词为 40 个，两词组配为 20 对。单个查询词又分为通用领域关键词和专业领域（经济类）关键词，均为 20 个，用于比较专业领域词汇和通用领域词汇用作查询词时，它们的相关词提示效果。封闭性测试数据和开放性测试数据的具体数量分布如表 1 所示。

表 1　相关性扩展功能测评数据来源一览表

类 别			数量
封闭性测试数据			20 对
开放测试数据	单个词	通用领域	20 个
		经济领域	20 个
	组配词		20 对

表 2　调研评估中用到的问题

问题号	评估标准
1	相关词平均个数
2	相关度
3	相关词提示的总体效果

3.1.2 选取调研对象

经过对 53 个搜索引擎的调研，选取有相关词提示功能的搜索引擎三个和全文数据库一个，分别为 Google[13]、百度[6]、雅虎[14]和 CNKI 全文数据库[15]。

3.1.3 “相关词”统计数据的获取

由于评价相关词提示效果的标准不确定，主观性强，不容易从定量角度对其进行测试。笔者针对每个开放性测试查询词，设计一个调查问题域，包括 3 个问题（见表 2）。其中问题 2 和 3 的评估，最好的方法是进行用户研究，它能证明搜索引擎/数据库提示的相关词是否真的能够帮助用户重新构造查询式，改善检索效果。将开放性测试查询词提交给搜索引擎和全文数据库进行信息检索，设计表格，保存系统返回的相关词。寻找三名同学作为志愿用户，对每一个相关词和其查询词/词组的相关度进行人工打分评估。人工打分规则如下：

4 分：相关词为查询词/词组的同义词或近义词；　**3 分**：相关词与查询词/词组非常相关；

2 分：相关词与查询词/词组比较相关；　**1 分**：相关词与查询词/词组有点相关；　**0 分**：相关词与查询词/词组不相关。

人工评分时，隐藏搜索引擎的名称，称为搜索引擎 A、B、C、D，以获得用户的无偏向反馈，统计每个查询词/词组在四个搜索引擎/数据库中与其相关词相关度的平均值和总值。最后，要求每个用户对相关词提示在改善检索效果方面作一个总体评价，评分范围是 0~10 分，总结为三个等级：好（7~10 分），一般（4~6 分），不好（0~3 分）。从而考察相关词提示在用户检索中重新构造查询式，改善检索结果的效果。

3.2 调研数据的分析

3.2.1 “查询扩展”分析

对“查询扩展”的分析，具体从两个角度进行：封闭性测试数据结果的分析和开放性测试数据的相关度人工评分的分析。具体分析过程是，横向比较不同属性的查询词，纵向比较四个搜索引擎/数据库。

（1）封闭性测试数据结果的分析

20 对（40 个）关键词，共获取 160 个数据，分为有相关性（每组中一个词在另一个词的相关词中）和没有相关性（不在相关词中）。经过初步统计，获取结果如表 3 所示。由表 3 可见，160 次相关性判断中仅有 46 次表

现为有相关性，占总数的 28.75%。说明搜索引擎提示的相关词覆盖率比较低，相当一部分有很高相关度的词没有被归纳到系统提示的相关词中。其中，考察每一对关键词，彼此出现在对方相关词中的比例更小，只有 6 对词同时在对方的相关词中。比较四个搜索引擎/数据库，封闭性测试中 Google 表现最好，相关词覆盖率为 52.5%，其次是百度。另外，6 对互相出现在对方相关词中的关键词中有 5 对来自于 Google，说明 Google 提供的相关词之间互引率高。

表 3　封闭性测试数据统计表

属性		百度	Google	CNKI	雅虎	总数
有相关性	个数	10	21	9	6	46
	百分比	25%	52.5%	22.5%	15%	28.75%
无相关性	个数	30	19	31	34	114
	百分比	75%	47.5%	77.5%	85%	71.25

（2）开放性测试数据的相关度人工评分分析

根据人工打分规则，三名用户在隐藏搜索引擎名称的情况下，将开放性测试数据作为查询式，对系统返回的相关词分别进行打分，统计结果如表 4 所示。由平均得分知，用户认为相关词与原查询词相关度不高。单个查询词平均得分为 1.5，介于“有点相关”和“比较相关”之间。组配查询词得分仅为 0.7，介于“不相关”与“有点相关”之间。因此，搜索引擎提示的相关词没有得到用户的认同，偏离了用户的检索意图，最终无法改善检索领域存在的信息过载问题。综合人工打分评估结果，四个搜索引擎的排名依次是：Google，百度，CNKI，雅虎。

表 4　相关度人工打分结果统计表

类别 / 结果	SE-A			SE-B			SE-C			SE-D			四个 SE 平均数		
	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
相关词个数	9	9	9	8.1	5.6	0.2	9.3	9	9	9.3	9.4	8.5	8.9	8.3	6.7
总得分	18.9	17.9	8.6	13.6	11.4	0.4	7.5	7.4	3.6	13.4	14.1	5.9	13.4	12.7	4.6
平均得 分	2.1	2.0	1.0	1.5	2.0	2.0	0.8	0.8	0.4	1.4	1.5	0.7	1.5	1.5	0.7

注：SE-A 、SE-B、SE-C、SE-D 分别为 Baidu、Google、Yahoo 中文、CNKI，用户评价时不知道具体的搜索引擎或全文数据库名称，下同。a、b、c 分别表示一般领域、经济领域、组配词等关键词类型。

3.2.2 “查询式专指度”分析

查询式的专指度，包括查询式所属学科领域的专业性程度，查询式表达概念的宽泛性和查询词个数的多少。根据人工打分统计表，发现一般领域和专业领域，如经济领域，查询关键词的相关词提示差别不大，相关词平均个数介于 8~9 之间，相关度平均得分均为 1.5。说明相关词提示与查询词所属领域的专业性关联不大。

比较单个查询词和组配查询词，发现相关词个数接近，但人工评分差距大，组配查询词相关度仅为 0.7，是单个查询词的 47%。单个查询词和组配查询词的主要区别，是专指度不相同，单个词的专指度低，组配词的专指度比较高。可以推测相关词提示效果与查询词的专指度成反比关系，即查询词专指度越高，相关词提示效果越差。

相关词包括同义词、近义词、上下位词等。实际调查中发现，当查询式专指度低时，返回的相关词中下位词多，用户给这类词打分高。说明用户在检索时希望提高查询式的专指度，而用户的查询式往往比较宽泛，因此给能够提高查询式专指度的下位词打分高。

相关词个数方面，Google 的区分度最好，不同属性查询式返回的相关词个数相差很大，单个词查询平均返回 7 个相关词，而组配词查询平均返回 0.2 个，其中有 17 个查询词没有相关词返回。百度、CNKI 和雅虎的人工评分结果，都反映出相关度和查询式专指度成反比例关系。由于在组配词查询中，Google 仅有 3 个查询词有相关词返回，而且每个查询词返回的相关词个数仅为 1~2 个，不能体现相关度与查询式专指度的关系。

3.2.3 相关词提示总体效果评价

综合上述两方面的评价，发现相关词提示在一定程度上能够改善检索效果，提高用户满意度。有研究显示，相关词提示在信息检索中被认为对检索成效帮助很大[16]。然而，相关词提示效果的评估牵涉到个人的主观判断，要评价相关词提示的效果，并不容易。

表 5 用户对相关词提示效果的综合评价

分 类	SE-A	SE-B	SE-C	SE-D	总体印象
一般领域查询词	一般	一般	不好	一般	一般
经济领域查询词	一般	一般	不好	一般	一般
组配词	不好	一般	不好	不好	不好

从用户体验角度出发，就相关词提示在改善检索方面的效果，进行分析评价。经过用户打分评价、作者整理统计，得到三名用户对相关词提示效果的综合评价表。由表 5 可以看出，用户对相关词提示效果的综合评价不高，其中单个查询词返回的相关词优于组配查询词返回的相关词。说明，搜索引擎提示的相关词与用户的检索意图相关度低，导致重新构造的查询式没有能够表达用户的信息需求。因此，要达到利用相关词提示来改善检索效果的目的，首先必须提高相关词与原查询词的相关度，使相关词更准确地表达用户的检索需求。与人工打分结果一致，用户对 Google 的相关词提示效果评价最高，说明 Google 在相关词提示方面的技术和效果均优于其他三个搜索引擎。四个搜索引擎/数据库相一致的是，单个查询词获取的相关词，在改善检索效果方面的作用，都优于组配查询词获取的相关词。

4 结束语

目前，搜索引擎和全文数据库提示的相关词还不能很好地表达用户的信息需求，没有达到改善检索效果的目的。但一项对 4000 名网络用户的调查表示：一次搜索失败后，有 76%的用户会重新构造查询式在同一个搜索引擎上再次进行信息检索（NPD2000）[17]。这意味着，很多用户依赖于搜索引擎系统来重新构造查询式，更好地表达他们的信息需求。因此，搜索引擎/数据库应该改进相关词提示的功能，在相关词中多加入查询词的同义词、近义词，提高相关词与原查询词的相关度。除了信息检索中的相关词提示，相关词的应用还可以推广到其他领域。例如电子商务中的 B2C 模式，顾客搜索一种商品，除了返回该商品的信息，还可以根据相关词原理推荐其他一些用户可能感兴趣的商品。

参考文献：

[1] http://www.google.com/intl/zh-CN/corporate/index.html, Accessed: 2006,3.1.

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

[3] Chien-Kang Huang, Lee-Feng Chien, Yen-Jen Oyang. Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs[J].Journal of American Society for Information Science and Technology,2003,54(7):638-649.

[4] 曾元显.关键词自动撷取技术与相关词回馈[J].中国图书馆学会会报,1997,59:59-64.

[5] Koenemann,J. Relevance feedback: usage, usability, utility[D]. Rutgers University, Dept. of Psychology,1996.

[6] http://www.baidu.com,Accessed: 2006,3.1.

[7] http://www.visimo.com , Accessed: 2006,3.1.

[8] 孙建军，成颖等编著. 信息检索技术[M].北京：科学出版社, 2004.

[9] http://blog.csdn.net/chinacommander/archive/2005/12/20/556981.aspx,Accessed: 2006,3.1.

[10] Beeferman, D. and Berger, A. Agglomerative clustering of a search engine query log[A]. In Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C], Boston: ACM Press, 2000: 407-415.

[11] Anick, P., Tiperneni, S. The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking[A]. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Berkeley, Boston: ACM Press,,1999:153-159.

[12] CSSCI. http://cssci.nju.edu.cn. Accessed: 2006,3.1.

[13] Google , http://www.google.com/webhp?hl=zh-CN . Accessed: 2006,3.1.

[14] 雅虎中文, http://www.yahoo.com.cn/. Accessed: 2006,3.1.

[15] 中国知网., http://www.edu.cnki.net/. Accessed: 2006,3.1.

[16] William B. Frakes and Ricardo Baeza-Yates, Information Retrieval: Data Structure and Algorithms[M], Prentice Hall, 1992.

[17] NPD Search and Portal Site Survey. Published by NPD New Media Services. (2000), http://www.searchenginewatch.com . Accessed: 2006,3.1