ISSN 1000-0054　清华大学学报（自然科学版）2005 年 第 45 卷 第 S1 期　28/39
CN 11-2223/N　J Tsinghua Univ (Sci & Tech), 2005, Vol. 45, No. S1　1857-1860

一种词汇共现算法及共现词对检索系统排序的影响

陈　翀，　彭　波，　闫宏飞，　王继民

（北京大学 信息科学技术学院，北京 100871）

摘　要：为了探讨共现词对检索系统排序相关性的影响，提出一种新的共现词汇算法——FDC。算法中考虑了词汇在文档中的共现频度、相对距离和共文档率。从天网搜索引擎查询日志中选取部分查询词，用本算法和潜在语义索引(LSI)方法分别求其共现词汇，并以相同的评分策略改变原始排序结果。Discounted cumulative gain (DCG)评估结果表明，本算法获得的共现词在 99%的置信度下对原始排序的相关性有改进；而 LSI 方法获得的共现词对排序相关性也表现出同样显著的改进效果。结果显示共现词汇能改进检索系统结果排序的相关性，并且不依赖于特定算法。

关键词：共现词汇；排序；相关性；信息检索

中图分类号：TP 391.1　　**文献标识码**：A

文章编号：1000-0054(2005)S1-1857-04

A term co-occurrence algorithm and the effect of co-occurrence terms on result ranking for information retrieval

CHEN Chong, PENG Bo, YAN Hongfei, WANG Jimin

(School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China)

Abstract: Terms which co-occur with query words are hypothesized to be helpful to discriminate the source documents. The effect of co-occurrence terms on reranking the relevant documents in information retrieval systems was studied in this paper. A new algorithm—FDC (frequency, term distance, co-collection ratio) is proposed to extract the most significant terms co-occurring with query words in documents. The algorithm considers both single document statistics (i.e., co-occurrence frequency, term distance) and global statistics in the collection (i.e., co-collection ratio). The performance of reranking is evaluated with discounted cumulative gain, based on the query and clickthrough logs of Tianwang search engine. Comparing the performance of FDC and the latency semantic indexing (LSI) method to extract co-occurrence terms, we found that retrieval performance with FDC reranking is improved from the baseline with a believability of 99%; similar result is got with LSI method. The result shows that the co-occurrence terms of query words can improve the relevance by reranking the original results, and does not depend on specific algorithms.

Key words: co-occurrence term; ranking; relevance; information retrieval

共现是指词汇在文档集中共同出现。以一个词为中心，可以找到一组经常与之搭配出现的词，作为它的共现词汇集，该集合描述了该词的语义上下文或语境。人们对共现词汇用于查询扩展的研究开展已久，但对其效果的评价并不一致。早期一个极端的结论认为共现词汇在查询扩展中的作用甚至不如随机选取的词汇[1]。文[2,3]分析认为文档集内容异构性、查询词的选取没有区分度有可能导致共现词查询扩展失败。

作者认为，共现词汇应用于检索，帮助提高返回结果的相关性可以从两方面考虑：一是查询扩展，二是对结果排序的影响。鉴于像搜索引擎这样的通用文档检索系统很难回避内容异构性，本文没有用共现词来扩展查询，而是用来筛选和调整相关结果排序，即研究对一个相关结果集，根据共现词在文中出现的情况调整文档的排序，使得结果的相关性排序能更好地体现用户的潜在需求，从而提高检索系统的排序质量。

Attar 和 Fraenkel[4]提出关联聚类和距离聚类均有助于寻找文档中的共现词汇。文[5]的研究也表明词汇的距离与共现紧密程度有关。除此之外，潜在语义索引 LSI 也可以用于求共现词汇[6]。参照上述研究本文提出了一种新的共现词汇算法——FDC (frequency, term distance, co-collection ratio)。

1　共现词汇算法 FDC

1.1　定　义

设文档集合 D，用 $|D|$ 表示集合 D 中元素个数，$|D|=n$，$d_j\in D$，$j\in[1,n]$；文档集所有词汇的集合为 T，$|T|=m$，$t_i\in T$，$i\in[1,m]$；关键词集合

收稿日期：2005-05-20

基金项目：国家自然科学基金重点资助项目（60435020）；
教育部博士点基金项目（20030001076）

作者简介：陈翀(1975-)，女(满)，辽宁，博士研究生。

通讯联系人：闫宏飞，讲师，E-mail：yhf@net.pku.edu.cn

为 K，取 $k\in K$；k 的共现词集合 $S_k\subset T$，且 $K\subset T$。

$\forall\ t_i\in T$ 且 $t_i\neq k$，t_i 在 d_j 中出现的词频为 f_{id_j}，如果不出现则 $f_{id_j}=0$。t_i 与 k 在 d_j 中共现频度为 $f_{id_j}\cdot f_{kd_j}$。t_i 与 k 在 d_j 中距离远近也反映了两者共同出现时的亲密关系，词间距离指两词间词汇的个数。因为在文档 d_j 中两个词均可能多次出现，故取它们在共现时的最短距离，用 $r_{(i\text{-}k)_{d_j}}$ 表示。如果 t_i、k 没有共同出现于 d_j，则 $r_{(i\text{-}k)_{d_j}}=\infty$，实际计算中将其设为构成 d_j 的所有词汇数。词汇的共现频度和距离体现了关键词及其共现词在同一文档中的关系。

t_i 相对于 k 的共文档率：在 k 出现的文档集 D_k 中同时出现 t_i 的文档数 $|D_k|\cap|D_{t_i}|$ 与 $|D_k|$ 的比。共文档率体现的是关键词及其共现词在整个文档集合中的关系。

1.2 FDC 共现词汇计算

用 $F_{i\text{-}k}$ 表示 t_i 与 k 在文档中的共现频度关系；$R_{i\text{-}k}$ 表示 t_i 与 k 在文档中的距离关系；$P_{i\text{-}k}$ 表示 t_i 与 k 共文档率。

$$F_{i\text{-}k}=\sum_{d_j\in D,j=1}^{n}(f_{id_j}\cdot f_{kd_j}),\tag{1}$$

$$R_{i\text{-}k}=n\Big/\sum_{d_j\in D,j=1}^{n}(1/r_{(i\text{-}k)_{d_j}}),\tag{2}$$

$$P_{i\text{-}k}=(|D_k|\cap|D_{t_i}|)/|D_k|.\tag{3}$$

定义 $C_{i\text{-}k}$ 表示在包含关键词 k 的所有文档中，t_i 和 k 共现的密切程度，即

$$C_{i\text{-}k}=\frac{F_{i\text{-}k}\cdot P_{i\text{-}k}}{R_{i\text{-}k}}.\tag{4}$$

对 $\forall\ k_l\in K$，$l\in[1,L]$，利用式(1)-(4)(实际计算中对式(1)作归一化[4])可以求出反映 T 中每个词与 k_l 共同出现的一系列 $C_{i\text{-}k_l}$，$i\in[1,m]$。排序取前 u 个元素，找到 T 中对应的词，就构成按照 FDC 方法所求的 k_l 的共现词汇集 S_{k_l}。

对于中文和英文的情况，细节处理有所不同，如果是中文，文档需要事先切分成词，词汇距离才可以求出。如果是英文，比较不同词的最短距离时，要转换大小写以便确认两个不同位置的词是否相同。

2 共现词典的建立

基于北大天网搜索引擎实际系统(http://e.pku.edu.cn)，从搜索引擎日志中选取部分查询词，在相关网页集中找到其共现词，建立以所选查询词为关键词条的共现词典，并进行实验分析。

2.1 建立步骤

依据局部聚类的思想来建立共现词典，共现词汇取自特定的文档集，也反映关键词在这个文档集中的上下文。合理选取用户最常用或最关心的词，建立它们的共现词典，可使大部分查询的检索结果得到优化。为此我们确定如下步骤建立共现词典：

1) 筛选关键词：从 2004 年 5 月 11 日到 2005 年 1 月 20 日之间的天网查询日志中找出现频率高于 100 的查询串，进行数据清理，作为预备关键词。

2) 构造包含关键词的文档集合：从天网返回关键词的前 30 个查询结果，每个网页(每个查询结果)的摘要文本构造为该关键词对应的一个文档，每个关键词对应 30 个文档。天网摘要是从网页中提取的包含关键词，长度不大于 180 字(360 个字节)的文本段。这种摘要方式符合共现词汇对词汇距离的关注，省去了对网页原文中大量远距词的处理。此外，搜索引擎返回的结果本身已经过内部评分排序，网页内容相对于关键词的相关性较好。这相当于变相操作了大量网页集合，并得到相关取向一致的样本文档，能够得到较好的效果保证。

3) 文本预处理：对每个文档切词，计算各词的位置、词频、共文档率。确定每个文档中关键词的位置，计算其他词的相对距离。

4) 计算：处理每个关键词对应的文档集 D，按前面式(1)-(4)计算 D 中各个词对关键词的共现情况。提取 $C_{i\text{-}k}(i\in[1,m])$ 最高的前 10 个词(不足 10 个词的取实际数目)，得到该关键词的共现词集合 S。

最终，用 FDC 方法建立了一个拥有 18988 个关键词条目、每个条目有不超过 10 个共现词汇的共现词典。

2.2 分 析

上述步骤 1 中，选取查询串时只提取了单个词作关键词，并且没有考虑关键词的词性。但从搜索引擎查询日志提出的有意义查询串分单个词和多词组合的情况，并且关键词有多种常用词性，通常名词表达较多的实际信息，对查询系统重要性较高，因此下一步工作可进一步考虑这些因素。高频查询词的文档区分能力不好[3]，相应的共现词汇可能五花八门，用在查询扩展中效果不好。但本研究探讨的是共现词对检索排序的影响，所以找关键词时并未回避高频查询词。后续可以尝试选取低频查询词，观察它们的共现词汇对检索排序是否有影响。

上述步骤 2 中，受天网对网页摘要长度的限制，不是原网页中所有出现关键词的周边文本段都被提取到构造的文档中，有些共现词可能被漏掉而影响共现词集合的质量，后续工作可以提取全部摘要再作尝试。

上述步骤 3 中，只对文档做切词，没有对关键词做切词。实际上，切词软件从语言的理解上可能会将文档中出现的关键词分为多个词。这会导致在切分后的词汇中找不到关键词，因而无法确定它的位置及其对各个词汇的距离。因此计算中一旦探测到这个情况即停止处理该关键词。

2.3　共现词典的应用

在所建立的共现词典中，多数情况下共现词集合能反映关键词的应用场景，并描述其语境上下文。这样的词典对于检索系统有如下的应用。

1）帮助指引用户确立自己查询意图。例如 $k=$远程教学，$S=$｛系统，网络，技术，信息，教育，管理，电视，大学｝，S 中的词能帮助了解“远程教学”的大致范畴，有助于用户分清自己要查询的具体领域如何界定；

2）帮助揣测用户意图、调节检索结果排序，将包含信息量大的文档位序提前。例如 $k=$戴佩妮，S 中包含歌手、下载、歌曲、mp3、专辑等，如果搜索引擎用户当前查询词为“戴佩妮”，潜在意图是希望找到一个可以下载戴佩妮歌曲的地方，系统返回结果前根据共现词再自动筛选一次，将包含上述词汇的网页位序提前，正好可以使用户的意图得到满足。

然而这种辨别意图的准确性如何，即通过共现词汇干预排序结果，是否真能符合用户对相关性的要求是我们关心的问题。为此通过分析用户点击情况的 DCG（discounted cumulative gain）[7] 实验，来评估共现词对检索排序的影响效果。

3　共现词典对检索排序效果的评估实验

3.1　实验设计

3.1.1　数据选择

建立共现词典用到的文档是取自天网在 2004 年 11 月间建立的网页索引数据，这批数据提供服务至 2005 年 2 月。我们采集高频查询词作为关键词，取网页摘要建立共现词典（包括用于对比实验的 LSI 词典）均使用这个批次的索引数据。评估实验采用的索引数据于 2005 年 4 月更新，提供服务至 2005 年 5 月 20 日，相应地采用这个期间的用户点击日志进行分析。

文档索引更新使同一个查询对应的结果集发生变化，即生成共现词典的网页集合与检验共现词典的网页集合不是同一个。因此，利用共现词影响检索结果排序时能够保证干预效果的真实性。由于天网是增量搜集网页，这两个集合中数据的最长搜集时间间隔为 6 个月，根据文[8]所述，同一时间存在的网页总体半衰期大约为 0.99 年，所以这段时间内虽会有一定数量的网页内容发生变化，整体信息变化不会太快。这意味着利用局部聚类的思想建立的共现词汇仍可以反映关键词的语义和上下文语境，因而可以产生有效的排序干预。

3.1.2　共现词影响检索排序的方式

通过检查和比较结果网页中包含共现词汇的个数，在已被系统判别为相关的原始结果集中，根据共现词的出现进一步调整相关性。

按照对用户检索行为的观察，实验中选择天网原始查询结果的前 20 条结果作重新排序处理。原因如下：1）前 20 个结果的摘要占据前 2 个返回页面，用户实际点击查看内容的网页在前两个结果页的可能性更大，而对于翻看更多结果页面往往没有耐心。2）用户希望搜索引擎能够快速返回结果，不希望漫长等待。文中对天网原始检索结果重新排序时，既使只选用简单的共现词出现数目做计算指标也会延误系统返回结果，所以只对前 20 条原始结果作重新排序。分析点击日志时也只考察一次查询中排在前 20 位以内的网页点击情况。

3.1.3　对比试验的准备

查询评估测试时间从 2005 年 5 月 1 日至 20 日，时间较短。为了避免结果分散，只选取 100 个关键词，用 FDC 和 LSI 两种方法在同样的文档集合上分别找到它们的共现词汇，考察它们对检索排序的影响，并与天网原始排序作对比。3 种排序情况分别用 FDC、LSI 和 Baseline 表示。

3.1.4　相关性评估

当用户提交一个查询词时，系统等概率随机选择上述 3 种排序结果，并记录用户点击查询结果的情况。在系统的用户点击日志中可以找到系统选择了哪种方法生成的词典、所点击文档在整个结果集中的排序、同一文档被不同查询点击的次数。这样每个查询词对应每种排序影响有一个用户点击的序列。

假设用户点击行为在统计意义上表明“用户认为点击对象与查询相关”，被点击次数高的网页相关性好，未点击的网页与查询无关[9]，那么通过比较这不同序列的差异，就可以得到不同排序的差异。利用点击日志评估的优点是用户完全不知道评估试验的存在，主观偏性小，试验可靠性好，并且没有人工评估开销；缺点是为了积累足够多在共现词典中出现的查询需要很长时间。

3.2　实验结果

统计每个关键词的查询结果点击情况，选取同

时拥有 FDC、LSI 和 Baseline 3 组数据且被点击次数足够大的关键词,一共获得 28 个关键词对应的结果序列。使用 DCG 作为评估指标。DCG 是用一个值表达多级相关度的指标,用查询结果点击次数的归一化数值作为序列〈$G[1],G[2],G[3]\cdots$〉的评分。DCG 的计算如下:

$$\mathrm{DCG}[i]=\begin{cases}G[1], & i=1;\\ \mathrm{DCG}[i-1]+G[i]/\log_b i, & i>1.\end{cases}$$

对 FDC、LSI 和 Baseline 了组数据,根据评估指标计算 DCG 分值并求平均,得到每个查询词 3 种情况下的 DCG 平均值分别为 0.641 93、0.619 09、和 0.428 29。

一般认为,用户对查询结果的点击情况反映了他对结果相关性的判断。但考虑到随机因素的存在,本研究利用统计显著性检验来分析不同排序方式给出的结果是否满足要求。

实验采用成对数据的 t 测试,置信度不低于 95%,通过测试,认为共现词汇对相关性排序存在影响,和天网原始排序相比存在差异。t 测试表明 FDC 组和 Baseline 组的对比结果能够通过参数为 0.01 的检验,即在置信度为 99%的情况下,FDC 算法所求的共现词汇影响原始检索结果排序更与用户查询意图相符。LSI 组数据与 Baseline 组对比,也表现出相似的结果。FDC 和 LSI 则没有显著差异,如图 1 所示。

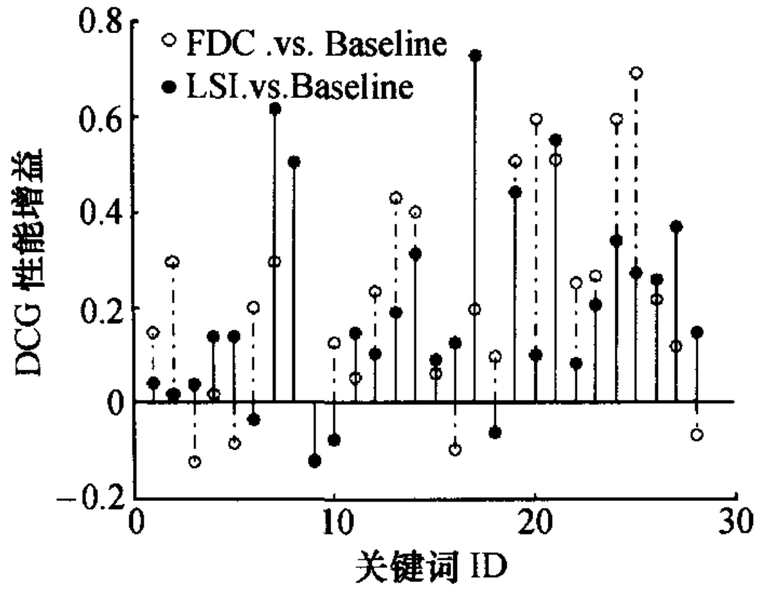

图 1 各关键词对应查询的 DCG 性能对比

在图 1 中用空心圆标记 FDC 组和 Baseline 组的对比结果,即对每个关键词的若干次查询中,FDC 共现词汇影响排序下用户对结果点击情况的 DCG 分值与原始排序下用户点击情况的 DCG 分值之差。用实心圆标记 LSI 组和 Baseline 组的对比结果,即 LSI 共现词汇对排序影响的 DCG 分值与原始排序下的 DCG 分值之差。图 1 中位于横轴以下的点数目远远小于横轴以上点的数目,说明以 DCG 评价来看,对关键词查询结果使用了 FDC 和 LSI 方法获得的共现词汇影响检索结果排序后,大多数情况对排序呈现显著的正向影响。

4 结论与展望

1) 提出一种新的计算共现词汇的算法 FDC,考虑词汇在文档中的共现频度、距离、共文档率等因素。

2) 利用 FDC 算法为天网搜索引擎建立了一个包含 18 988 个关键词条目、可以用于实际检索系统的共现词典。

3) 发现用查询词的共现词集合重新调整原始检索结果的排序,就能够使新的排序对相关性上的改进获得 99%的置信度。

4) 对比 LSI 方法获得的共现词汇在同样条件下对排序的影响,发现改进效果也同样存在,即共现词对检索系统排序结果相关性的影响并非依赖于特定算法。这一结论有助于探讨共现词汇对于改进海量文档规模的检索系统的结果排序的一般性意义。

该研究验证了 FDC 方法得到的共现词汇对与检索排序相关性有正面影响,但在何种影响算法或因子下才可得到最优结果,使共现词汇对排序相关性的正面影响最大化,有待于进一步研究。

致谢 本文工作全过程得到了李晓明老师及北京大学网络与分布式实验室多位同学的支持,在此向他们表示感谢。

参考文献 (References)

[1] Smeaton A F, Rijsbergen C J V. The retrieval effects of query expansion on a feedback document retrieval system [J]. *Computer Journal*, 1983, **26**(3): 239 - 246.

[2] Peat H J, Peter W. The limitations of term co-occurrence data for query expansion in document retrieval systems [J]. *Journal of the American Society for Information Science*, 1991, **42**(5): 378 - 383.

[3] Efthimiadis E N. Query expansion [A]. Williams, Martha E, eds. Annual Review of Information Systems and Technology [C]. 1996: 121 - 187.

[4] Attar R, Fraenkel A S. Local feedback in full-text retrieval systems [J]. *J ACM*, 1977, **24**(3): 397 - 417.

[5] Beefenman D, Berger A, Laferty J. A model of lexical attraction and repulsion [A]. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics [C]. 1997.

[6] Berry M W, Dumais S T, O'Brien G W. Using linear algebra for intelligent information retrieval [J]. *SIAM Review*, 1995, **37**(5): 573 - 595.

[7] Järvelin Kalervo, Kekäläinen Jaana. IR evaluation methods for retrieving highly relevant documents [A]. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. 2000: 41 - 48.

[8] 李晓明. 对中国曾有过静态网页的一种估计 [J]. 北京大学学报(自然科学版), 2003, **39**(5): 394 - 398.
LI Xiaoming. An estimate on Chinese historical static web pages [J]. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 2003, **39**(5): 394 - 398. (in Chinese)

[9] 彭波. 搜索引擎检索系统的效率优化和效果评估研究 [D]. 北京: 北京大学, 2004.
PENG Bo. On Efficiency Optimization and Effectiveness Evaluation of Search Engine Retrieval System [D]. Beijing: Peking University, 2004. (in Chinese)