

搜索引擎页面排序算法研究综述*

李绍华, 高文宇

(广东商学院 信息学院, 广东 广州 510320)

摘要: 系统地分析了现有的页面排序算法, 指出了它们各自的优势和存在的不足, 并指出不同算法在不同领域和场合所具有的优势。建立专业搜索引擎是提高搜索准确性和性能的有效途径。通过网格技术将各种专业搜索引擎集成在一起, 形成一个基于网格的搜索引擎, 从而更好地满足不同背景不同偏好的用户需求。

关键词: 搜索引擎; 页面排序; 链接分析

中图分类号: TP393.09

文献标志码: A

文章编号: 1001-3695(2007)06-0004-04

Survey of Page-ranking Algorithms

LI Shao-hua, GAO Wen-yu

(School of Information Science, Guangdong University of Business Studies, Guangzhou Guangdong 510320, China)

Abstract: Several page-ranking algorithms were analyzed systematically. In conclusion, different page-ranking algorithms had different performance in different domains; thus, developing domain specific algorithms was a effective method to improve performance of search engine. Moreover, through grid technique, it integrated several domain specific algorithms and formed a grid-based search engine.

Key words: search engine; page-ranking; link analysis

随着 Internet 的飞速发展, 其提供的文档(网页)也以惊人的速度在增长。有关的调查统计表明, Internet 上的网页每不到一年的时间就会增长一倍。要从这么大量的信息库中提取出有用的信息就越来越依赖于搜索引擎的功能。而网页的排序则是搜索引擎要解决的关键问题之一。

Sergey Brin 等人^[1]提出 PageRank 算法开启了链接分析研究的热潮。基于链接分析的算法, 提供了一种衡量网页质量的客观方法; 独立于语言, 独立于内容; 无需人工干预就能自动发现 Web 上的重要资源, 挖掘出 Web 上的重要社区, 自动实现文档分类。PageRank 在 Google 中的应用获得了巨大的商业成功。在最初的 Google 中, 首先使用 IR(Information Retrieve) 算法找到所有与查询关键字相匹配的网页; 然后根据页面因素(标题、关键字密度等)进行排名; 最后通过 PageRank 得分调整网站排名结果。

近几年来, 基于链接分析的页面排序算法一直是一个热点问题, 学者提出了许多页面排序算法。

1 PageRank 及其相关算法

基于链接分析的排序算法中, 最为著名的就是 PageRank。所谓链接分析主要基于如下两个重要假设: 超文本链接包含了用户对一个网站的判断信息; 对一个网站而言, 如果其他网站链接到该网站的入链数越多, 该网站越重要。以上假设在各种基于链接分析的算法中均以某种方式体现出来。

1.1 PageRank 算法

PageRank 算法是最早提出的链接分析算法之一, 并被

Google 用于计算网页的重要性得分。其基本思想是: 如果网页 T 存在一个指向网页 A 的链接, 则表明 T 的所有者认为 A 比较重要, 从而把 T 的一部分重要性得分赋予 A 。这个重要性得分的值则由 T 的 PageRank 值 $PR(T)$ 和 T 的出链(从 T 链出的链接)数 $C(T)$ 决定。具体公式为: $PR(T) / C(T)$ 。而对于页面 A 其 PageRank 值 $PR(A)$ 的计算如下:

$$PR(A) = PR(T_1) / C(T_1) + \dots + PR(T_n) / C(T_n) \quad (1)$$

其中, T_1, T_2, \dots, T_n 为含有指向 A 链接的页面。

为了避免 Link Sink(许多网页没有入链或出链)问题, 对式(1)引入一个阻尼系数 d , 使其变为

$$PR(A) = (1 - d) + d[PR(T_1) / C(T_1) + \dots + PR(T_n) / C(T_n)] \quad (2)$$

如此经过多次迭代, 系统的 PR 值达到收敛。

PR 的计算公式可以从概率的角度解释为一个随机网络冲浪者随机选择一个网页后, 不断地点击网页上的链接, 但是从不返回; 除非最后厌烦了才随机选择另一个页面。随机冲浪者访问某个页面的随机概率就是该页面的 PageRank 值; 阻尼系数 d 就是随机冲浪者在某个页面会厌烦然后选择一个新页面的概率。页面的 PageRank 值越高, 则随机冲浪者发现它的概率亦越高。这种思路非常富有创意。一个网页的外部链接越多, 则对网络冲浪者来说, 发现它的机会也就越大。

文献[2]结合近年来 Web 出现的一些新特性对 PageRank 提出了一些改进措施。文献[3]中对 PageRank 算法中的阻尼系数 d 进行了深入讨论, 从理论上分析了 d 的取值不同对于 PageRank 算法效果的影响。文献[4]提出了一种方法用于对 PageRank 中的迭代计算进行加速。

PageRank 的一个优势在于它是一个与查询无关的静态算法, 因此所有网页的 PageRank 值均可以通过离线计算获得。这样有效地减少了在线查询时的运算量, 极大地降低了查询响应时间。

然而 Internet 上的内容涵盖了众多主题, 在现实应用中, 人们的查询所希望得到的信息往往是具有某一方面主题特征的, 而 PageRank 仅仅依靠计算网页的外部链接数量来决定该网页的排名, 而忽略了页面的主题相关性, 从而影响了搜索结果的相关性和准确性。另一方面, PageRank 算法对新网页有很严重的歧视性, 因为一个新网页入链数量通常都很少, 自然 PR 值很低。

1.2 Topic-Sensitive PageRank

由于 Internet 上的内容千差万别, 涵盖众多不同的领域和主题。同样一个查询如“汽车”, 可能用户 1 是想买一台汽车, 他感兴趣的是汽车品牌、价格; 而用户 2 是想参加与汽车相关的运动, 他感兴趣的是与汽车相关的运动项目和赛事。因此要想给用户返回更为准确的查询信息就有必要基于不同的主题来对页面排序。最初的 PageRank 算法中是没有考虑主题相关因素的。主题敏感 PageRank 算法 (Topic-Sensitive PageRank, TSPR)^[5] 正是在这种背景下提出来的。

TSPR 核心思想就是通过离线计算, 计算出一个 PageRank 向量集合 (在 PageRank 算法中, 仅计算一个 PageRank 向量), 该集合中的每一个向量与某一主题相关, 即计算某个页面关于不同主题的得分。例如某个网页在教育这个主题的得分为 a , 在体育这个主题的得分为 b, \dots 。

具体来说, TSPR 也可分为两个主要阶段:

(1) 主题相关的 PageRank 向量集合的计算。先将所有页面的内容划分为 16 个主题, 根据 Crawler 搜集来的网页, 计算该网页在不同主题的得分情况, 即不同的 PageRank 向量。

(2) 在线查询, 主题的确定。在线查询阶段, 先根据用户的搜索请求确定用户的 Context (用 q 表示); 然后使用式 (3) 计算用户的 Context 属于不同主题 c_j 的概率 $P(c_j | q)$; 最后使用式 (4) 计算网页的综合得分 S_{qd} , 并根据该得分进行页面排序。式 (4) 中的 rank_{jd} 即页面 d 在主题 c_j 的得分情况。

$$P(c_j | q) = P(c_j) \times P(q | c_j) / P(q) = P(c_j) \times \prod_i P(q_i | c_j) \quad (3)$$

$$S_{qd} = \sum_j P(c_j | q) \times \text{rank}_{jd} \quad (4)$$

根据用户的查询请求和相关 Context 判断用户查询相关的主题 (即用户的兴趣取向), 从而提高返回结果的准确性无疑是一种有效的方法。

遗憾的是 TSPR 并没有利用主题的相关性来提高链接得分的准确性。事实上对于网页类别的划分可以更有效地计算链接的价值和权威性。例如评阅论文时, 经常需要填写对相关领域的熟悉程度。也就是说, 评阅者对论文所属的领域越熟悉, 则评阅者所给出的评分越可信, 从而在最后的计算中拥有更高的权重。

对于网页之间的链接分析与上述论文评阅的例子类似。可以把网页 A 指向网页 B 的链接视为 A 对 B 的评分; 若 A 与 B 的内容是相近的, 则 A 的评分更为可信。例如一个教育相关的网站 A 指向另一个教育相关的网站 B, 较一个娱乐相关的网

站 C 指向教育相关的网站 B 更为权威、可信。

因此, 可以将上述思想应用到 PageRank 的 PR 值计算中。这将在今后的研究工作中作进一步的考虑。

1.3 Hilltop

Hilltop^[6] 算法的指导思想与 PageRank 是一致的, 即通过链接的数量和质量来确定搜索结果的排序权重。与 PageRank 不同的是, 在 Hilltop 中仅考虑那些专家页面 (Expert Sources), 即专门用于引导人们浏览资源的页面。Hilltop 在收到一个查询请求时, 首先根据查询的主题计算出一列相关性最强的专家页面, 然后根据指向目标页面的非从属专家页面的数量和相关性来对目标页面进行排序。目标页面的排序得分反映了与查询主题相关的最好的独立专家页面的集体意见。若在此过程中, Hilltop 无法得到一个足够大的专家页面集合, 则返回空值。Hilltop 算法主要包含两个步骤:

(1) 专家页面搜索。所谓专家页面就是关于某个主题的包含着很多非从属页面链接的网页。非从属页面是指两个页面分别属于两个来自非从属组织的作者。在预处理阶段, 由搜索引擎的 Crawler 搜集来的网页的一个子集被辨识为专家页面集。

辨识专家页面的关键主要有: 剔除从属页面; 选择专家页面 (Out-Link 大于阈值 k); 对专家页面进行索引。

当收到一个查询时, 从专家页面集中挑选出与查询主题相关的专家页面子集。

(2) 目标页面排序。Hilltop 算法认为“一个目标页面在某个查询主题是权威的当且仅当有一些与该查询主题相关的最好的专家页面指向该目标页面。”

然而, Hilltop 在应用中还存在如下一些问题:

专家页面的搜索和确定对算法起关键作用, 专家页面的质量决定了算法的准确性; 而专家页面的质量和公平性在一定程度上难以保证。同时 Hilltop 忽略了大多数非专家页面的影响。在 Hilltop 的原型系统中, 专家页面只占到整个页面的 1.79% (2.5 ~ 140 M), 在一定程度上并不能很好地反映整个 Internet 的民意。

Hilltop 算法在无法得到足够的专家页面子集时 (小于两个专家页面), 返回为空, 即 Hilltop 适合于对查询排序进行求精, 而不能覆盖。这意味着 Hilltop 可以与某个页面排序算法结合, 提高精度, 而不适合作为一个独立的页面排序算法。Hilltop 中根据查询主题从专家页面集合中选取与主题相关的子集也是在线运行的, 这与前面提到的 HITS 算法一样会影响查询响应时间。随着专家页面集合的增大, 算法的可伸缩性存在不足之处。

2 HITS 及其相关算法

2.1 HITS 算法

HITS (Hypertext-Induced Topic Search) 算法是 Kleinberg^[7] 提出的。它是 IBM Almaden Research Center 的“CLEVER”研究项目的一部分。

对于每个页面 P , HITS 算法计算两个值, 即 Authority 和 Hub。Authority 和 Hub 之间满足如下关系: 对于 Authority, 如果

有越多具有好 Hub 的页面指向 P , P 的 Authority 值就越大; 对于 Hub, 如果 P 指向越多具有好 Authority 的页面, P 的 Hub 值就越大。对整个 Web 集合而言, Authority 和 Hub 是相互依赖、相互加强的关系。Authority 和 Hub 之间相互优化的关系, 即为 HITS 算法的基础。

在 HITS 算法中, 将查询 q 提交给传统的基于关键字匹配的搜索引擎。搜索引擎返回很多网页, 从中取前 n 个网页作为根集 (Root Set), 用 S 表示。 S 满足三个条件: S 中网页数量相对较小; S 中网页大多数是与查询 q 相关的网页; S 中网页包含较多的权威网页。

通过向 S 中加入被 S 引用的网页和引用 S 的网页, 将 S 扩展成一个更大的集合 T 。

以 T 中的 Hub 网页为顶点集 V_1 , 以权威网页为顶点集 V_2 , V_1 中的网页到 V_2 中的网页的超链接为边集 E , 形成一个二分有向图 $SG = (V_1, V_2, E)$ 。对 V_1 中的任一个顶点 v , 用 $h(v)$ 表示网页 v 的 Hub 值; 对 V_2 中的顶点 u 用 $a(u)$ 表示网页的 Authority 值。开始时 $h(v) = a(u) = 1$, 对 u 执行 I 操作修改其 $a(u)$, 对 v 执行 O 操作修改其 $h(v)$; 然后规范化 $a(u)$ 、 $h(v)$; 如此不断重复计算下面的操作 I、O, 直到 $a(u)$ 、 $h(v)$ 收敛。HITS 算法输出一组具有较大 Hub 值的网页和具有较大权威值的网页。

实验数据表明, HITS 的排名准确性要比 PageRank 高。但是 HITS 最大的问题在于它是一个依赖于查询关键字的算法, 在线运算量大, 极大地影响了算法的可伸缩性, 从而难以应用于大规模的网页数据集。

HITS 算法还存在着主题漂移问题, 即紧密链接 TKC (Tightly-Knit Community Effect) 现象。如果在集合 T 中有少数与查询主题无关的网页, 但是紧密链接的, HITS 算法的结果可能就是这些网页。因为 HITS 只能发现主社区, 偏离了原来的查询主题。

用 HITS 进行窄主题查询时, 可能产生主题泛化问题, 即扩展以后引入了比原来主题更重要的新主题, 新主题可能与原始查询无关。泛化的原因是因为网页中包含不同主题的向外链接, 而且新主题的链接更加具有重要性。

针对 HITS 的这些问题提出了许多改进算法, 如 SALSA^[8]、BFS^[9]、PHITS^[10] 等。

2.2 SALSA

PageRank 算法是基于用户随机的向前浏览网页的直觉知识, HITS 算法考虑的是 Authority 网页和 Hub 网页之间的加强关系。实际应用中, 用户大多数情况下是向前浏览网页, 但是很多时候也会回退浏览网页。基于上述直觉知识, R. Lempel 和 S. Moran 提出了 SALSA (Stochastic Approach for Link-Structure Analysis) 算法。该算法考虑了用户回退浏览网页的情况, 保留了 PageRank 的随机漫游和 HITS 中把网页分为 Authority 和 Hub 的思想, 取消了 Authority 与 Hub 之间的相互加强关系。

具体算法如下:

(1) 与 HITS 算法的第一步一样, 得到根集并且扩展为网页集合 T , 并除去孤立节点。

(2) 从集合 T 构造无向图 $G = (V_h, V_a, E)$:

$$V_h = \{ S_h \mid S \in C \text{ and out-degree}(S) > 0 \} \text{ (} G \text{ 的 Hub 边)}$$

$$V_a = \{ S_a \mid S \in C \text{ and in-degree}(S) > 0 \} \text{ (} G \text{ 的 Authority 边)}$$

$$E = \{ (S_h, r_a) \mid S_h \rightarrow r_a \text{ in } T \}$$

这就定义了两条链: Authority 链和 Hub 链。

(3) 定义两条马尔可夫链的变化矩阵, 也是随机矩阵, 分别是 Hub 矩阵 H 和 Authority 矩阵 A 。

$$H_{i,j} = \frac{1}{|F(i)|} \sum_{k \in F(i)} \frac{1}{|B(k)|} \times 1$$

$$A_{i,j} = \frac{1}{|B(i)|} \sum_{k \in B(i)} \frac{1}{|F(k)|} \times 1$$

(4) 求出矩阵 H 和 A 的主特征向量, 得到对应的马尔可夫链的静态分布。

(5) A 中值大者对应的网页就是所要找的重要网页。

SALSA 算法没有 HITS 中相互加强的迭代过程, 计算量远小于 HITS。SALSA 算法只考虑直接相邻的网页对自身 AH 的影响; 而 HITS 是计算整个网页集合 T 对自身 AH 的影响。

试验结果表明, HITS 算法结果集中于主题的某个方面。而 SALSA 算法的结果覆盖了多个方面, 也就是说, 对于 TKC 现象, SALSA 算法比 HITS 算法有更高的健壮性。

2.3 BFS

SALSA 算法计算网页的 Authority 值时, 只考虑网页在直接相邻网页集中的受欢迎程度, 忽略了其他网页对它的影响。HITS 算法考虑的是整个图的结构, 特别地, 经过 n 步以后, 网页 i 的 Authority 的权重是 $|BF^n(i)| / |BF^n|$ 。 $BF^n(i)$ 为离开网页 i 的 $(BF)^n$ 的路径数目, 即网页 $j \rightarrow i$, 对 i 的权值贡献等于从 i 到 j 的 $(BF)^n$ 路径数量。如果从 i 到 j 包含有一个回路, 那么 j 对 i 的贡献将会呈指数级增加, 这并不是算法所希望的, 因为回路可能不是与查询相关的。

Allan Borodin 等人提出了 BFS (Backward Forward Step) 算法, 既是 SALSA 的扩展情况, 也是 HITS 的限制情况。其基本思想是, SALSA 只考虑直接相邻网页的影响, BFS 扩展到考虑路径长度为 n 的相邻网页的影响。在 BFS 中, $BF^n(i)$ 被指定表示能通过 $(BF)^n$ 路径到达 i 的节点集合, 这样 j 对 i 的贡献就依赖于 j 到 i 的距离。BFS 采用指数级降低权值的方式, 节点 i 的权值计算如下:

$$a_i = 2^{n-1} |B(i)| + 2^{n-2} |BF(i)| + 2^{n-3} |BFB(i)| + \dots + |BF^n(i)|$$

算法从节点 i 开始, 第一步向后访问, 然后继续向前或向后访问邻居; 每一步遇到新的节点加入权值计算, 节点只有在第一次被访问时加入进去计算。

2.4 PHITS

D. Cohn and H. Chang 提出了计算 Hub 和 Authority 的统计算法 PHITS (Probabilistic Analogue of the HITS)。他们提出了一个概率模型。在这个模型中, 一个潜在的因子或主题 z 影响了文档 d 到 c 的一个链接。PHITS 算法进一步假定, 给定因子 z , 文档 c 的条件分布 $P(c|z)$ 存在, 并且给定文档 d , 因子 z 的条件分布 $P(z|d)$ 也存在。

$$P(d, c) = P(d) \times P(c|d)$$

其中, $P(c|d) = \sum_z P(c|z) \times P(z|d)$ 。根据这些条件分布, 提出了一个可能性函数 $L: L = \sum_{(d,c)} P(d, c)$ 。 M 是对应的连接矩阵。

PHITS 算法使用 Dempster 等人^[11] 提出的 EM 算法分配未知的条件概率, 使得 L 最大化, 即最好地解释了网页之间的链

接关系。算法要求因子 z 的数目事先给定。Allan Borodin 等人^[9]指出, PHITS 中使用的 EM 算法可能会收敛于局部最大化, 而不是真正的全局最大化。D. Cohn 等人^[12]还提出了结合文档内容和超链接的概率模型。

3 页面排序算法的一些新观点

3.1 Link Fusion

鉴于目前大多数页面排序算法只分析包含在 Web 页面中的链接, 文献[13]提出了 Link Fusion 页面排序算法。在该算法中, 将链接分为两类: Intra-type Links。用于表示同一数据空间中的数据对象关系, 多指包含在 Web 页面中的链接。

Inter-type Links。用于表示不同数据空间中数据对象之间的关系, 多指用户、查询条件与 Web 页面之间的关系。在链接分析中, 同时考虑了 Intra-type Link 和 Inter-type Link 的影响。

具体来说, 用户和他们提交的查询条件以及用户浏览的 Web 页面分别代表三个数据空间。当用户提交查询请求时、用户浏览 Web 页面时、一个查询参考其他 Web 页面时, 这三个不同的数据空间便被联系起来。三种操作 (Submit、Browse、Reference) 包含了这三个不同数据空间之间的 Inter-type Link。因此在进行页面排序时, 应该不仅仅考虑 Intra-type Link, 还要考虑浏览 Web 页面的用户以及参考这些 Web 页面的查询请求。

3.2 确定用户的特性和目标——CubeSVD

为了提高用户查询结果的准确性, 一些算法通过用户的查询日志 (Query Log) 来确定用户的偏好, 进而找出用户的目的。这是非常有效的方法。例如前面提到的 TSPR 就是希望确定用户的主题, 从而能更准确地返回查询结果。在文献[14]中, 提出了一种新的用于确定用户目标的方法——User-click Behavior。它是利用用户点击数据 (Clickthrough) 来提高搜索引擎的效果。一个搜索引擎每天都要接受大量的查询请求, 将用户提交的查询请求以及用户所点击的查询结果页面记录下来, 然后通过对这些 Clickthrough 数据的分析, 获得用户的兴趣以及用户定位信息资源的模式, 从而更准确地执行用户的查询请求。

4 结束语

Internet 上信息量的爆炸式增长使得人们越来越依赖于搜索引擎获取所需的信息。虽然目前的商用搜索引擎获得了巨大的成功, 但其中还有许多方面可以进一步完善。本文通过对现有搜索引擎页面排序算法的分析, 希望为今后的工作提供一些基础性支持。

AT&T 香农实验室的 Brian Amento 指出, 用权威性来评价网页的质量与人类专家评价的结果是一致的, 并且各种链接分析算法的结果在大多数情况下差别很小^[15]。通过前面对现有页面排序算法的分析也可以看出, 不同算法在不同领域和场合有各自的优势。对 Internet 来说, 页面有着众多不同的主题 (领域), 用户有着各种各样的背景和偏好, 因此难以用一种页面排序技术来满足所有的需求。今后对于搜索引擎的研究可以着眼于建立一些专业搜索引擎, 即针对不同应用场合

和应用领域建立不同用途的专业搜索引擎; 然后利用网格技术, 建立基于网格的搜索引擎体系结构, 将各种不同的专业搜索引擎联合起来, 结合对用户背景和偏好的自动分析, 自动引用不同的专业搜索引擎, 从而为用户提供更为精确的搜索结果。

参考文献:

- [1] RIN S, PAGE L. The anatomy of a large-scale hyper textual Web search engine: proc. of the 7th International World Wide Web Conference [C]. [S. l.]: [s. n.], 1998.
- [2] EIRON N, MCCURLEY K S. Link analysis: ranking the Web frontier: proc. of the 13th Conference on World Wide Web [C]. [S. l.]: [s. n.], 2004.
- [3] BOLDI P, SANTINI M, VIGNA S. PageRank as a function of the damping factor: proc. of the 14th Conference on World Wide Web [C]. [S. l.]: [s. n.], 2005.
- [4] MCSHERRY F. A uniform approach to accelerated PageRank computation: proc. of the 14th Conference on World Wide Web [C]. [S. l.]: [s. n.], 2005.
- [5] HAVELIWALA T H. Topic-sensitive PageRank: proc. of the 11th International World Wide Web Conference [C]. [S. l.]: [s. n.], 2002.
- [6] KRISHNA B, GEORGE A M. When experts agree: using non-affiliated experts to rank popular topics: proc. of the 10th International World Wide Web Conference [C]. [S. l.]: [s. n.], 2001.
- [7] KLEINBERG J M. Authoritative sources in a hyper-linked environment [J]. Journal of the ACM, 1999, 46 (5): 604-632.
- [8] LEMPEL R, MORAN S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect: porc. of the 9th International World Wide Web Conference [C]. [S. l.]: [s. n.], 2000.
- [9] BORODIN A, ROBERTS G O, ROSENTHAL J S. Finding authorities and hubs from link structures on the world wide Web: proc. of the 10th International WWW Conference [C]. [S. l.]: [s. n.], 2001.
- [10] COHN D, CHANG H. Learning to probabilistically identify authoritative documents: proc. of the 17th International Conference on Machine Learning [C]. [S. l.]: [s. n.], 2000.
- [11] DEMPSTER A, LAIRD N, RUBIN D. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society: Series B, 1977, 39 (1): 1-38.
- [12] COHN D, HOFMANN T. The missing link: a probabilistic model of document content and hypertext connectivity: advances in Neural Information Processing Systems [C]. [S. l.]: [s. n.], 2000.
- [13] XI Wensi, ZHANG Benyu, CHEN Zheng, *et al.* Link fusion: a unified link analysis framework for multi-type interrelated data objects: proc. of the 13th International WWW Conference [C]. [S. l.]: [s. n.], 2004: 319-327.
- [14] SUN Jiantao, ZENG Huajun, LIU Huan, *et al.* CubeSVD: a novel approach to personalized Web search: proc. of the 14th International WWW Conference [C]. [S. l.]: [s. n.], 2005: 382-390.
- [15] BRIAN A, LOREN T, WILL H. Does authority mean quality? predicting expert quality ratings of Web documents: proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. [S. l.]: [s. n.], 2000.