元搜索引擎系统合成算法的约束条件*

阳小华[1], 刘振宇[1], 谭敏生[1], 刘 杰[1], 张敏捷[2]

[1](南华大学 计算机科学与技术学院,湖南 衡阳 421001);
[2](澳大利亚伍龙贡大学 信息技术与计算机科学学院,澳大利亚)
E-mail: xiaohua1963@yahoo.com.cn
http://www.nhu.edu.cn

摘要: 合成是元搜索引擎系统中一个重要的技术问题.给出了搜索引擎和元搜索引擎的形式化定义,对各种可能的元搜索引擎合成类型进行了划分.在此基础上,提出了元搜索引擎合成的一般性约束条件以及针对特殊类型的特殊约束条件.这些约束条件为构造好的元搜索引擎合成策略提供了基本保障.

关 键 词: 因特网;搜索引擎;元搜索引擎;排序;合成
中图法分类号: TP393 文献标识码: A

搜索引擎(search engine,简称 SE)是一种基于关键字检索的因特网信息查询工具,其核心是一个排序系统.由于因特网上的信息量巨大,没有一个搜索引擎能够覆盖整个网络.为了获得所需的信息,人们有时不得不使用多个搜索引擎.而元搜索引擎(meta search engine,简称 MSE)则是多个搜索引擎的集成,其工作过程可以归纳为如下 6 步:(1) 接受用户的原始查询;(2) 把原始查询分别转换为各个成员搜索引擎能够接受的形式;(3) 向成员搜索引擎发送查询;(4) 收集各个搜索引擎的原始查询结果;(5) 对原始查询结果进行合成,形成最终结果;(6) 把最终查询结果递交给用户.Metasearch, SavvySearch,Metacrawler,Profusion,Inquirus 和 MetaGer 等是一些有代表性的元搜索引擎系统.

从元搜索引擎的工作原理可以看出,查询结果合成是一个十分重要的环节.由于搜索引擎查询结果的规模常常比较庞大,而用户又通常缺乏足够的耐心、精力和时间去遍历所有的命中文档,他们一般只会检查前几条或几十条信息,因此,最终查询结果中各个项目的排列顺序是至关重要的.

查询结果合成在分布式信息检索系统中得到了广泛的关注,人们提出了许多种合成方法.在文献[1]中,J.P. Callan 等人针对不同的情况给出了 4 种典型的合成算法.(1) 如果只有文档的原始顺序是已知的,则可以采用间隔排列合成法:首先把每个查询结果中的第 1 项交叉列出,然后再把各个查询结果中的第 2 项交叉列出,依此类推.(2) 如果可以得到文档的原始相关性分值,那么当这些分值可以直接比较时,则可以采用原始分值合成法:直接依据每个文档的原始相关性分值决定其合成排列次序.(3) 如果文档的原始分值不能直接比较,则可以通过对 idf(倒排文档频率)等进行标准化来得到规范的相关性分值,并以之为根据确定文档的合成排列次序(规范分值合成).(4) 加权分值法:首先计算出各个信息源相应于查询条件的重要性,再以此为权乘上文档的相关性分值作为决定其合成排列次序的根据.

在文献[2]中,Kirsch 给出了另一类典型的合成方法.这种方法要求对下层搜索引擎进行一些修改,以便返回

* 收稿日期: 2000-09-13; 修改日期: 2001-03-12
基金项目: 国家留学基金委资助项目(97517007);湖南省教育厅科研基金资助项目(00C174);澳大利亚纽卡索尔大学 ERC 研究基金资助项目(142/1078);南华大学博士启动基金资助项目(5-01-XJQ-001)

作者简介: 阳小华(1963－),男,湖南衡阳人,博士,教授,主要研究领域为信息检索技术,数据库技术,工作流管理;刘振宇(1962－),男,湖南衡阳人,副教授,主要研究领域为信息检索技术;谭敏生(1965－),男,湖南衡阳人,副教授,主要研究领域为数据库技术;刘杰(1974－),男,湖南衡阳人,助教,主要研究领域为数据库技术;张敏捷(1956－),女,吉林长春人,高级讲师,主要研究领域为专家系统,agent 技术.

诸如各个搜索项在文档中出现的次数和在整个数据库中出现的次数等额外信息.元搜索引擎则利用这些信息在客户端重新计算文档的相关性,并依此为根据决定文档的最终排列顺序.

实际的元搜索引擎系统所使用的合成方法是各种各样的.Metacrawler[3]引入概念可信度来决定文档与查询的相关程度.Metacrawler 把可信度的取值范围限定在 0~100 0 之间.每个搜索引擎查询结果中第 1 项的可信度初值为 1 000,第 2 项的可信度初值为 999,依次递减.重复出现文档的可信度等于其所有初值之和.从本质上看,Metacrawler 的合成策略允许各个搜索引擎就最终结果中文档的排列顺序进行投票,被多个搜索引擎选中的文档更有可能排在只被一个搜索引擎选中的文档前面.

Profusion[4]的合成算法其实就是规范分值法和加权分值法的一种集成.它由3个步骤组成:首先把搜索引擎给出的文档与查询之间的(原始)相关值规范映射到[0,1],然后把规范分值乘上搜索引擎的权,最后如果有重复出现的文档,则取其中的最大值作为文档的(最终)相关值.SavvySearch[5]的合成方法就是规范分值法,重复出现的文档则以其相关性分值之和作为排序的依据.

Inquirus[6]采用了客户端重新计算文档相关性的合成策略.与 Kirsch 不同,Inquirus 首先对搜索引擎查询结果中的文档进行下载,然后再在客户端独立计算文档与查询之间的相关性.

虽然合成问题在 MSE 系统和其他分布式信息检索系统中得到了广泛的关注,但是,有一个关于合成的基本问题却从来没有被明确地讨论过即一个合理的合成算法需要满足的必要条件或基本限制.在分布式专家系统中,合成的类型、策略和约束等问题已经得到了较好的研究[7].而在分布式信息检索系统中却没有一个明确的算法合理性概念,也没有一个能够从理论上保证算法合理性的准则.

本文将集中讨论一个合理的 MSE 合成策略必须满足的约束条件.第 1 节给出 MSE 合成问题的形式化定义.在第 3 节中,将给出合成策略必须满足的一般性约束条件和在特定情况下应该满足的特殊约束条件.最后总结全文并给出今后工作的简单展望.

1 问题的提出

定义 **1**. 搜索引擎 SE 是一个四元组,$SE=\langle D,Q,r,t\rangle$,其中 D 是 Internet 文档索引数据库,Q 是查询条件集,r 是排序算法,$t(0<t)$是查询结果选择标准.

对于一个 Internet 文档,不同的搜索引擎通常有不同形式的索引.但是本文将忽略索引的具体格式,而把它简单地等同于文档对象本身.不同的搜索引擎能够接受的查询条件集通常也是不一样的.在本文中,我们也不考虑查询条件的具体表示形式.一般来说,r 是一个从 $D\times Q$ 到 $R^{+}\cup\{0\}$ 的映射,其中 R^{+} 是正实数集.我们通常把 $r(d,q)$ 理解为文档 d 与查询条件 q 的相关性.

对于一个给定的查询条件q,搜索引擎产生查询结果的过程可以简单地理解为3个步骤:第1,对于D中的每一个文档d,计算出它与q的相关性$r(d,q)$;第2,根据相关性$r(d,q)$对文档集合D进行排序;第3,选取所有满足条件$t\le r(d,q)$的相关文档组成查询结果.查询结果最常见的表示形式是由命中的文档按相关性以从大到小为序组成的序列.

定义**2**. 元搜索引擎MSE是一个五元组,$MSE=\langle E_m,Q_m,H_m,r_m,t_m\rangle$,其中$E_m$是成员搜索引擎的集合,$Q_m$是查询条件集,$H_m$是查询条件变换集,$r_m$是排序方法,$t_m(0<t_m)$是查询结果选择标准.

假设 $E_m=\{SE_1,SE_2,...,SE_n\}$,$H_m=\{h_1,h_2,...,h_n\}$,则对任意的 q 属于 Q_m,$SE_1,SE_2,...,SE_n$ 所接受的查询分别为 $h_1(q)$,$h_2(q),...,h_n(q)$.假设 $SE_i(1\le i\le n)$产生的原始查询结果集合为 $R_m(q)=\{d_i^1,d_i^2,...,d_i^{k(i)}\}(1\le i\le n)$,$MSE$ 的最终查询结果集合为 $R_i(q)=\{d^1,d^2,...,d^k\}$,则下面的条件成立:

(1) $R_i(q)\subseteq R_1(q)\cup...\cup R_n(q)$;

(2) 对任意的 $1\le i\le k$,$t_m\le r_m(d^i,q)$;

(3) 对任意的 $1\le i,j\le k$,如果 $O_m(d^i)\le O_m(d^j)$,则 $r_m(d^i,q)\le r_m(d^j,q)$,其中 $O_m(x)$表示 x 在合成结果序列中的位置.

定义**3**. 假设$MSE=\langle E_m,Q_m,H_m,r_m,t_m\rangle$是一个元搜索引擎,$SE_i=\langle D_i,Q_i,r_i,t_i\rangle(1\le i\le n)$是$MSE$的成员搜索引擎,$f_{\Delta,\theta}:P^n\to P$($P$是排序算法的集合)称为$MSE$的结果合成映射,如果$r_m=f_{\Delta,\theta}(r_1,r_2,\ldots,r_n)$.

对任意的 q 属于 Q_m,MSE 的最终查询结果可以理解为一个 k 维的向量 $V_m(q)=(w_m^1,w_m^2,...,w_m^k)$,其中

$w_m^j=r_m(d^j,q)(1\le j\le k)$.类似地,$SE_i(1\le i\le n)$产生的原始查询结果也可以对应于一个 k 维向量 $V_i(q)=(w_i^1,w_i^2,...,w_i^k)$,其中 $w_i^j=r_i(d^j,q)$,当 d^j 属于 $R_m(q)$时;否则,d^j=UNKNOWN.因此 *MSE* 查询结果合成的基本任务就是在 $V_i(q)(1\le i\le n)$的基础上构造出 $V_m(q)$.

由于 $V_i(q)(1\le i\le n)$的一些元素是未知的,而且来源于不同搜索引擎的文档原始相关性通常也是不能直接比较的,因此,从概念上看,*MSE* 中的合成通常由两个步骤组成:首先,把原始查询结果向量 $V_i(q)(1\le i\le n)$修剪为一种规范的形式$\Delta V_i(q)=(\Delta w_i^1,\Delta w_i^2,...,\Delta w_i^k)$,其中$\Delta:R^+\cup\{0,\text{UNKNOWN}\}\to R^+\cup\{0\}$称为修剪映射;然后再把多个规范向量$\Delta V_i(q)(1\le i\le n)$归并为一个向量 $V_m(q)=\Delta V_1(q)\theta\Delta V_2(q)\theta\ldots\theta\Delta V_n(q)$,其中 $\theta:(R^+\cup\{0\})\times(R^+\{0\})\to R^+\cup\{0\}$称为归并映射.因此,对任意的 q 属于 Q_m,有

$$r_m(d^j,q)=f_{\Delta,\theta}(r_1,r_2,...,r_n)(d^i,q)=\Delta r_1(d^i,q)\theta\Delta r_2(d^i,q)\theta\ldots\theta\Delta r_n(d^i,q).$$

例1:(1) MetaSearch结果合成的修剪映射就是可信度的计算方法,而归并映射则是累加求和函数.

(2) 在 Profusion 中,结果合成的修剪映射由两个步骤组成:首先把搜索引擎给出的文档与查询之间的(原始)相关性分值规范映射到[0,1],然后把规范分值乘上搜索引擎的权;而归并映射则是求最大值函数.

(3) 在 SavvySearch 中,结果合成的修剪映射是相关性分值规范映射,而归并映射则是累加求和函数.

(4) Inquirus 结果合成的修剪映射是客户端的相关性独立计算函数,而归并映射则是求最大值函数.

2 查询结果合成的基本类型

根据 *MSE* 中各个搜索引擎的查询结果集合之间的关系,我们可以把 *MSE* 的查询结果合成划分为 4 种基本类型:对等、包含、不相交和交搭.

定义 **4**. 在一个 *MSE* 中,对于给定的查询条件 q,如果任意两个搜索引擎的查询结果集合都是相同的,则称 *MSE* 关于 q 的查询结果合成为对等合成.

定义 **5**. 在一个 *MSE* 中,对于给定的查询条件 q,如果存在一个搜索引擎,其他搜索引擎的查询结果集合都是它的查询结果集合的子集,并且至少有一个搜索引擎的查询结果集合是它的查询结果集合的真子集,则称 *MSE* 关于 q 的查询结果合成为包含合成.

定义 **6**. 在一个 *MSE* 中,对于给定的查询条件 q,如果任意两个搜索引擎的查询结果集合都是不相交的,则称 *MSE* 关于 q 的查询结果合成为不相交合成.

定义 **7**. 在一个 *MSE* 中,对于给定的查询条件 q,如果查询结果合成既不是对等的,也不是包含的或不相交的,则称 *MSE* 关于 q 的查询结果合成为交搭合成.

根据 *MSE* 中各个搜索引擎的查询结果序列中文档之间的次序关系,可以把 *MSE* 的查询结果合成划分为另外两种基本类型:相容与冲突.

定义 **8**. 在一个 *MSE* 中,对于给定的查询条件 q,假设 $SE_i=\langle D_i,Q_i,r_i,t_i\rangle$和 $SE_j=\langle D_j,Q_j,r_j,t_j\rangle$是任意两个成员搜索引擎,$R_i(q)$和 $R_j(q)$分别是 SE_i 和 SE_j 的查询结果集合,如果对任意两个同时属于 $R_i(q)$和 $R_j(q)$的 x,y, $r_i(x,q)\le r_i(y,q)$当且仅当 $r_j(x,q)\le r_j(y,q)$,则称 *MSE* 关于 q 的查询结果合成为相容合成;反之,则称为冲突合成.

因为不相交合成一定是相容合成,所以综合起来,*MSE* 查询结果合成有 7 种基本类型,它们是对等相容合成、对等冲突合成、包含相容合成、包含冲突合成、交搭相容合成、交搭冲突合成和不相交合成.

3 合成策略的约束条件

3.1 一般性约束条件

显而易见,*MSE* 的最终查询结果应该只与原始查询结果有关,而与合成的次序无关.所以,有以下的查询结果合成一般性约束条件:

(1) 结合性:$(\Delta r_i\theta\Delta r_j)\theta\Delta r_k=\Delta r_i\theta(\Delta r_j\theta\Delta r_k)$,其中 r_i,r_j,r_k 是成员搜索引擎的排序算法;

(2) 交换性: $\Delta r_i\theta\Delta r_j=\Delta r_j\theta\Delta r_i$,其中 r_i,r_j 是成员搜索引擎的排序算法.

因为查询结果合成必须满足结合性和交换性,所以在下面的讨论中,我们只针对两个搜索引擎的查询结果合成来讨论相应的约束条件.不失一般性,假设 $SE_i=\langle D_i,Q_i,r_i,t_i\rangle$和 $SE_j=\langle D_j,Q_j,r_j,t_j\rangle$是元搜索引擎 $MSE=\langle E_m,$

$Q_m,F_m,r_m,t_m\rangle$中的两个搜索引擎,q 是用户给出的查询条件,$R_i(q)$和 $R_j(q)$分别是 SE_i 和 SE_j 的查询结果集合,$R(q)=R_i(q)\cup R_j(q)$是合成结果文档集合.

因为元搜索引擎是在其他搜索引擎的基础上运行的,它通常没有自己的数据库,即使有,相对于搜索引擎的数据库,元搜索引擎的数据库也是较小的.因此在决定文档的最终排列次序时,元搜索引擎应该尊重其成员搜索引擎的共同意见.只有当成员搜索引擎的意见不一致时,元搜索引擎才能依据自己的知识作出裁决.如果某些文档之间的次序关系在所有的成员搜索引擎中都是一致的,那么经过元搜索引擎的合成以后,这些关系应该仍然成立.基于这种认识,我们得到了查询结果合成的第 3 个一般性约束条件:

(3) 公共次序不变性:

(a) $\forall x$ 属于 $R_i(q)\cap R_j(q)$,y 属于 $R_i(q)\cap R_j(q)$,若 $r_i(x,q)\le r_i(y,q)$并且 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$;

(b) $\forall x$ 属于 $R_j(q)-R_i(q)$,y 属于 $R_i(q)\cap R_j(q)$,若 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$;

(c) $\forall x$ 属于 $R_i(q)-R_j(q)$,y 属于 $R_i(q)\cap R_j(q)$,若 $r_i(x,q)\le r_i(y,q)$,则 $r_m(x,q)\le r_m(y,q)$.

条件(a)是显然合理的.条件(b)的合理性可以解析如下:$\forall x$ 属于 $R_j(q)-R_i(q)$,如果 x 属于 D_i,则 $r_i(x,q)\le t_i\le r_i(y,q)$,所以应该有 $r_m(x,q)\le r_m(y,q)$;如果 x 不属于 D_i,则 x 与其他文档之间的次序关系完全由 SE_j 决定,SE_i 不能发表任何意见,所以仍然有 $r_m(x,q)\le r_m(y,q)$.类似可得条件(c)的合理性.

如果约定 UNKNOWN 小于所有的非负实数,则可以得到下面的结果.

定理1. 假设$f_{\Delta,\theta}(r_1,r_2,\ldots,r_n)=\Delta r_1\theta\Delta r_2\theta\ldots\theta\Delta r_n$是$MSE$的合成算法,如果$f_{\Delta,\theta}$的修剪映射$\Delta$和归并映射$\theta$都是单调递增函数,则$f_{\Delta,\theta}$满足公共次序不变性.

例2: (1) MetaSearch合成算法的修剪映射和归并映射都是单调递增函数,所以它满足公共次序不变性.

(2) Profusion 合成算法的修剪映射和归并映射都是单调递增函数,所以它满足公共次序不变性.

(3) SavvySearch 合成算法的修剪映射和归并映射都是单调递增函数,所以它满足公共次序不变性.

(4) Inquirus 合成算法的归并映射是单调递增函数,但其修剪映射不能保证单调递增,所以它不能满足公共次序不变性.

结合性、交换性和公共次序不变性是每个合理的合成策略必须满足的约束条件.在不同的合成类型下,基于公共次序不变性的基本思想就可以得到相应的特定约束条件,为此首先引入两个新的概念.

定义9. 假设$MSE=\langle E_m,Q_m,H_m,r_m,t_m\rangle$是一个元搜索引擎,对于任意的$SE_i=\langle D_i,Q_i,r_i,t_i\rangle$和$SE_j=\langle D_j,Q_j,r_j,t_j\rangle$属于$E_m$,$q$属于$Q_m$,假设$R_i(q)$和$R_j(q)$分别是$SE_i$和$SE_j$的查询结果集合.

(1) 如果对任意的 x,y 属于 $D_i\cap D_j$,$r_i(x,q)\le r_i(y,q)$当且仅当 $r_j(x,q)\le r_j(y,q)$,则称 SE_i 和 SE_j 的排序算法在 MSE 中是一致的.

(2) 如果对任意的 x 属于 $R_i(q)$,若 x 属于 D_j,那么 x 属于 $R_j(q)$,反之亦然,则称 SE_i 和 SE_j 的结果选取标准在 MSE 中是一致的.

3.2 对等合成的约束条件

当MSE关于q的查询结果合成为对等合成时,各个搜索引擎查询结果的文档集合都是一样的,基于公共次序不变性,我们有下面的对等合成约束条件:

对于任意的x和y属于$R(q)$,如果$r_i(x,q)\le r_i(y,q)$并且$r_j(x,q)\le r_j(y,q)$,则$r_m(x,q)\le r_m(y,q)$.

3.3 包含合成的约束条件

当MSE关于q的查询结果合成为包含合成时,合成策略要满足的约束条件需根据搜索引擎之间的关系来定.不失一般性,假设$R_i(q)\subset R_j(q)$,则$R(q)=R_i(q)\cup R_j(q)=R_j(q)$.下面我们分3种情况进行讨论.

3.3.1 SE_i 和 SE_j 具有一致的排序算法

假设 x 和 y 是集合 $R_i(q)(R_i(q)=R_i(q)\cap R_j(q))$中的任意两个文档,如果 $r_j(x,q)\le r_j(y,q)$,则因为 SE_i 和 SE_j 具有一致的排序算法,所以一定有 $r_i(x,q)\le r_i(y,q)$,因而应该有 $r_m(x,q)\le r_m(y,q)$.

假设 x 或 y 是集合 $R_j(q)-R_i(q)$中的文档,如果 $r_j(x,q)\le r_j(y,q)$,则因为 SE_i 和 SE_j 具有一致的排序算法,所以当 x 和 y 都属于 SE_i 的索引数据库 D_i 时,仍然一定有 $r_i(x,q)\le r_i(y,q)$;若 x 或 y 不属于 D_i,那么 SE_i 对 x 和 y 之间的顺序

关系没有发言权,只能服从 SE_j 的意见.因而也应该有 $r_m(x,q)\le r_m(y,q)$.

综上所述,有以下包含合成约束条件:对于任意的 x 和 y 属于 $R(q)$,如果 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$.

3.3.2 SE_i 和 SE_j 具有一致的结果选取标准

假设x是集合$R_j(q)-R_i(q)$中的一个元素,因为SE_i和SE_j具有一致的结果选取标准,所以SE_i的索引数据库中一定没有x的索引信息,那么x和其他文档之间的次序只能由SE_j单独决定.而集合$R_i(q)$中文档之间的次序则必须由SE_i和SE_j共同决定.因此,当SE_i和SE_j具有一致的排序算法时,有以下包含合成约束条件:

(1) 对于任意的 x 或 y 属于 $R_j(q)-R_i(q)$,如果 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$;

(2) 对于任意的 x 和 y 属于 $R_i(q)$,如果 $r_j(x,q)\le r_j(y,q)$并且 $r_i(x,q)\le r_i(y,q)$,则 $r_m(x,q)\le r_m(y,q)$.

3.3.3 SE_i 和 SE_j 的排序算法和结果选取标准都是不同的

假设x是集合$R_j(q)-R_i(q)$中的文档,y是集合$R_i(q)$中的文档,当x属于SE_i的索引数据库D_i时,则由于x不属于$R_i(q)$而y属于$R_i(q)$,因此一定有$r_i(x,q)<t_i\le r_i(y,q)$;若x不属于D_i,那么SE_i对x和y之间的顺序关系没有发言权,只能服从SE_j的意见.所以,如果$r_j(x,q)\le r_j(y,q)$,则$r_m(x,q)\le r_m(y,q)$.各个搜索引擎共同的文档顺序关系应该保持不变.所以,有以下包含合成约束条件:

(1) 对于任意的 x 属于集合 $R_j(q)-R_i(q)$,y 属于集合 $R_i(q)$,如果 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$.

(2) 对于任意的 x 和 y 属于 $R_i(q)$,如果 $r_i(x,q)\le r_i(y,q)$并且 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$.

3.4 不相交合成的约束条件

3.4.1 SE_i 和 SE_j 具有一致的排序算法

假设x和y是$R_j(q)$中的任意两个文档,如果$r_j(x,q)\le r_j(y,q)$,那么:(a) 若x和y都属于D_i,则因为SE_i和SE_j具有一致的排序算法,所以一定有$r_i(x,q)\le r_i(y,q)$,因而有$r_m(x,q)\le r_m(y,q)$;(b) 若x或y不属于D_i,那么SE_i对x和y之间的顺序关系没有发言权,只能服从SE_j的意见,因而也应该有$r_m(x,q)\le r_m(y,q)$.

假设x和y是$R_i(q)$中的任意两个文档,如果$r_i(x,q)\le r_i(y,q)$,同理,应该有$r_m(x,q)\le r_m(y,q)$.

综上所述,当SE_i和SE_j具有一致的排序算法时,有以下不相交合成约束条件:

(1) 对于任意的 x 和 y 属于 $R_i(q)$,如果 $r_i(x,q)\le r_i(y,q)$,则 $r_m(x,q)\le r_m(y,q)$;

(2) 对于任意的 x 和 y 属于 $R_j(q)$,如果 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$.

3.4.2 SE_i 和 SE_j 具有一致的结果选取标准

假设x是$R_j(q)$中的一个元素,因为SE_i和SE_j具有一致的结果选取标准,而$R_j(q)\cap R_i(q)=\varnothing$,所以$SE_i$的索引数据库中一定没有$x$的索引信息,那么$x$和其他文档之间的次序只能由$SE_j$单独决定.同理,集合$R_i(q)$中文档之间的次序也只能由$SE_i$决定.因此,当$SE_i$和$SE_j$具有一致的排序算法时,有以下包含合成约束条件:

(1) 对于任意的 x 和 y 属于 $R_i(q)$,如果 $r_i(x,q)\le r_i(y,q)$,则 $r_m(x,q)\le r_m(y,q)$.

(2) 对于任意的 x 和 y 属于 $R_j(q)$,如果 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$.

3.5 交搭合成的约束条件

3.5.1 SE_i 和 SE_j 具有一致的排序算法

假设 x 和 y 是 $R_i(q)$中的两个文档,如果 $r_i(x,q)\le r_i(y,q)$,那么:(a) 若 x 和 y 同时属于 SE_j 的索引数据库,由于 SE_i 和 SE_j 具有一致的排序算法,因此一定有 $r_j(x,q)\le r_j(y,q)$;(b) 若 x 或 y 不属于 SE_i 的索引数据库,则 SE_i 对 x 和 y 之间的顺序排列没有发言权.无论哪种情况都应该有 $r_m(x,q)\le r_m(y,q)$.

同理,假设 x 和 y 是 $R_i(q)$中的两个文档,如果 $r_j(x,q)\le r_j(y,q)$,那么应该有 $r_m(x,q)\le r_m(y,q)$.所以当 SE_i 和 SE_j 具有一致的排序算法时,我们有如下的交搭合成约束条件:

(1) 对任意的 x 和 y 属于 $R_i(q)$,若 $r_i(x,q)\le r_i(y,q)$,则 $r_m(x,q)\le r_m(y,q)$;

(2) 对任意的 x 和 y 属于 $R_j(q)$,若 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$.

3.5.2 SE_i 和 SE_j 具有一致的结果选取标准

由于SE_i和SE_j具有一致的结果选取标准,若$R_i(q)$中的文档x不属于$R_j(q)$,则x不属于D_j,因此只属于$R_i(q)$而不属于$R_j(q)$的文档之间的相对次序只能由SE_i单独决定.同理,只属于$R_j(q)$而不属于$R_i(q)$的文档之间的相对次序只能

由SE_j单独决定.所以有以下交搭合成约束条件:

(1) 对于任意的 x 和 y 属于 $R_i(q)$,如果 x 或 y 属于 $R_i(q)-R_j(q)$并且 $r_i(x,q)\le r_i(y,q)$,则 $r_m(x,q)\le r_m(y,q)$;

(2) 对于任意的 x 和 y 属于 $R_j(q)$,如果 x 或 y 属于 $R_j(q)-R_i(q)$并且 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$;

(3) 对于任意的 x 和 y 属于 $R_j(q)\cap R_i(q)$,若 $r_i(x,q)\le r_i(y,q)$且 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$.

3.5.3 SE_i和 SE_j的排序算法和结果选取标准都是不一致的

假设x是集合$R_j(q)-R_i(q)$中的文档,y是集合$R_j(q)\cap R_i(q)$中的文档,如果$r_j(x,q)\le r_j(y,q)$,若x属于SE_i的索引数据库D_i时,则由于x不属于$R_i(q)$而y属于$R_i(q)$,因此一定有$r_i(x,q)<t_i\le r_i(y,q)$;若x不属于D_i,那么SE_i对x和y之间的顺序关系没有发言权,只能服从SE_j的意见.所以应该有$r_m(x,q)\le r_m(y,q)$.同理,当x是集合$R_i(q)-R_j(q)$中的文档时,我们可以得到类似的结论.另外,各个搜索引擎共同的文档顺序关系应该保持不变,所以有以下交搭合成约束条件:

(1) 对于任意的 x 属于 $R_j(q)-R_i(q)$,y 属于 $R_j(q)\cap R_i(q)$,若 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$;

(2) 对于任意的 x 属于 $R_i(q)-R_j(q)$,y 属于 $R_j(q)\cap R_i(q)$,若 $r_i(x,q)\le r_i(y,q)$,则 $r_m(x,q)\le r_m(y,q)$;

(3) 对于任意的 x 和 y 属于 $R_j(q)\cap R_i(q)$,若 $r_i(x,q)\le r_i(y,q)$且 $r_j(x,q)\le r_j(y,q)$,则 $r_m(x,q)\le r_m(y,q)$.

4 结 论

本文讨论元搜索引擎系统合成策略的约束条件.我们给出了搜索引擎和元搜索引擎的形式化定义,对各种可能的合成类型进行了划分.在此基础上,提出了元搜索引擎合成的一般性约束条件和针对特殊类型的特殊约束条件,它们是保证合成策略合理性的基础.遗憾的是,由于这个问题长期以来一直被忽略,因而有些合成算法不能完全满足必要的约束条件,其合理性是没有保障的.

从概念上看,一个合理的合成算法应该包含两个步骤:首先根据本文给出的约束条件决定文档之间的基本顺序关系;然后再利用其他方法确定遗留的次序关系.合成算法的构造策略将是我们今后要做的研究工作.

References:

[1] Callan, J.P., Lu, Z., Croft, W.B. Searching distributed collections with inference networks. In: Fox, E.A., Ingwersen, P., Fidel, R., eds. Proceedings of the 18th International Conference on Research and Development in Information Retrieval. ACM Press, 1995. 21~28.

[2] Kirsch, S.T. Document retrieval over networks wherein ranking and relevance scores are computed at the client for multiple database documents. United States Patent #5,659,732, 1997.

[3] Selberg, E.W. Towards comprehensive web search [Ph.D. Thesis]. University of Washington, 1999.

[4] Gauch, S., Wang, G., Gomez, M. Profusion: intelligent fusion from multiple, distributed search engines. Journal of Universal Computer Science, 1996,2(9):637~649.

[5] Lorence, S., Giles, C.L. Inquirus, the NECI meta search engine. Computer Networks and ISDN Systems, 1998,(30):95~105.

[6] Howe, A.E., Dreilinger, D. SavvySearch: a meta-search engine that learns which search engine to query. ACM Transactions on Information Systems, 1997,3(15):195~222.

[7] Zhang, M., Zhang, C. Potential cases, methodologies, and strategies of synthesis of solutions in distributed expert systems. IEEE Transactions on Knowledge and Database Engineering, 1999,3(11):498~503.

Constraints for Fusion Algorithms in Meta Search Engine Systems*

YANG Xiao-hua[1], LIU Zhen-yu[1], TAN Min-sheng[1], LIU Jie[1], ZHANG Min-jie[2]

[1](*School of Computer Science and Technology*, *South-China University*, *Hengyang* 421001, *China*);

[2](*School of Computer Science and Information Technology*, *University of Wollongong*, *Australia*)

E-mail: xiaohua1963@yahoo.com.cn

http://www.nhu.edu.cn

Abstract: Fusion is one of the main technical problems for running a meta search engine. In this paper, the formal definitions of search engine and meta search engine are given and the potential cases of fusion in meta search engine systems are classified. Based on this classification, the necessary constraints of fusion methods in different fusion cases are described. This provides a reliable foundation for the construction of good fusion algorithms in meta search engine systems.

Key words: Internet; search engine; meta search engine; rank; fusion

* Received September 13, 2000; accepted March 12, 2001

Supported by the China Scholarship Council under Grant No.97517007; the Scientific Research Found of Hu'nan Provincial Education Department under Grant No.00C174; the ERC Research Grant of the university of Newcastle of Australia under Grant No.142/1078; the Doctoral Research Foundation of South China University of China under Grant No.5-01-XJQ-001

第 9 届 Rough 集、模糊集、数据挖掘与粒度计算国际学术会议(RSFDGrC 2003)

征文通知

RSFDGrC 2003 是 RSFDGrC 系列国际会议的第九届会议，该系列会议每两年举行一次。RSFDGrC 2003 是该系列会议第一次在中国举办。RSFDGrC 2003 将由国家自然科学基金会、重庆邮电学院等有关单位共同主办，于 2003 年 5 月 26 日~29 日在重庆召开。会议将针对 Rough 集、模糊集、数据挖掘、粒度计算等计算智能的理论与应用进行学术研究与讨论。

一、 征文范围

Rough 集理论及应用	计算智能	机器学习	文字计算
Fuzzy 集理论及应用	粒度计算	软计算及其应用	进化计算
Petri 网	软计算的逻辑基础	非经典逻辑	神经网络
软计算复杂性	空间推理	统计推理	智能 Agent
多标准决策分析	决策支持系统	知识发现与数据挖掘	多 Agent 技术
网络智能	集成智能系统	近似推理与不确定推理	数据仓库
模式识别与图像处理	其他有关领域		

二、 征文要求

(1) 论文未被其他会议、期刊录用或发表；

(2) 大会工作语言为英语；

(3) 来稿一式 4 份；

(4) 为联系方便，请务必提供作者的姓名、单位、通信地址、电话、传真及 EMAIL 地址；

(5) 论文版面格式及投稿信息请参看会议主页。

三、 重要日期

特邀小组会议申请：2002 年 11 月 10 日（收到日期）

征文截止日期：2002 年 12 月 10 日（收到日期）

录用通知日期：2003 年 1 月 10 日（发出日期）

提交正式论文截止日期：2003 年 2 月 10 日（收到日期）

四、 联系方式

联系地址：400065 重庆市南岸区黄桷垭镇堡上园 1 号 重庆邮电学院计算机科学与技术研究所

联 系 人：RSFDGrC 2003 程序委员会 王国胤教授

联系电话：023-62460066 传真：023-62461882

E-mail：wanggy@cqupt.edu.cn Http://www.cqupt.edu.cn/rsfdgrc