一种基于用户偏好分析的查询优化方法

梅 翔　孟祥武　陈俊亮　徐 萌
(北京邮电大学网络与交换技术国家重点实验室　北京　100876)

摘　要：该文对如何满足不同兴趣用户的查询需求进行了研究，提出了一种基于用户偏好分析的查询优化方法。该方法将用户对网页的偏好转化为对本体知识库中实例的偏好；分析本体实例之间的语义关联，发现隐含的用户偏好；综合用户偏好历史，建立用户当前状态下偏好的数学模型，以预测用户对网页的关注程度。实现了相应的原型系统，实验结果表明，相对于传统的个性化搜索技术，该文提出的方法能更有效地获取用户偏好，适应用户偏好的变化，提高搜索引擎查询的准确率。

关键词：网络搜索；个性化；本体；语义关联
中图分类号：TP393　　**文献标识码**：A　　**文章编号**：1009-5896(2008)01-0033-05

A Query Optimization Method Based on User Profiles Reasoning

Mei Xiang　Meng Xiang-wu　Chen Jun-Liang　Xu Meng
(*State Key Laboratory of Networking and Switching Technology*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*)

Abstract: User profiles, descriptions of user interests, can be used by search engines to provide personalized search results. A query optimization method based on user profiles reasoning is presented. This method creates user profiles by classifying users' interests into instances in an ontology knowledge base, and then propagates user preferences to find users' latent interests by analyzing the semantic association among the ontology instances. It integrates users' current and history preferences to process the search results. A prototype system is implemented and the experimental results show that users' latent preferences can be learned accurately and personalized search based on user preference yields significant improvements over the origin results.

Key words: Web search; Personalization; Ontology; Semantic association

1　引言

上世纪 90 年代兴起的搜索引擎技术在一定程度上解决了 Web 信息检索困难的问题，然而传统的搜索引擎仅依据用户输入的查询关键字和网页的重要性，判定网页与用户查询需求的相关度，忽略了用户个体的特点。使用相同的关键字，背景和兴趣爱好不同的用户查询的对象可能完全不同，如房地产代理用关键字“office price”查询办公场所的出租行情，IT 采购员用同样的关键字查询微软出品办公软件的价格。个性化搜索技术[1]旨在提供满足用户个性化需求的搜索服务，通过分析用户的偏好，预测用户对网页的关注程度。目前存在许多个性化搜索系统，GroupLens[2]要求用户给出对网页的显式评价，通过用户之间的相似性分析预测用户对网页的关注情况，该方法完全依赖用户显式的反馈获取用户偏好，时间开销较大，且无法预测用户对新出现网页的兴趣。文献[3,4]分析用户浏览的网页文本，将词语作为记录用户兴趣的基本单位，但方法中词语只是孤立的逻辑符号，忽略了词语之间的内在联系。文献[5,6]将网页归类到具有特定意义的概念，用概念权重来表示用户偏好，但该方法在分析用户偏好时，忽略了用户兴趣随时间变化的特性和概念之间的联系。

本文针对现有个性化搜索方法存在的问题，提出了一种基于用户偏好分析的查询优化方法 (Query Optimization Method based on User profiles Reasoning，QOMUR)。该方法能够显式和隐式地获取用户兴趣，通过网页归类，将用户对网页的偏好转化为对本体知识库中实例的偏好，借助本体实例之间语义关联的推理，实现用户偏好的扩散，可以发现隐含的用户偏好。本文综合时间因素，建立用户偏好的数学模型，以预测用户对网页的关注程度，并实现了相应的原型系统 (Semantic Assistant System for Personal Web Search，SASPWS)，该系统运行于客户端，对传统搜索引擎返回的检索结果进行优化。实验表明，本文提出的方法可以准确全面地获取用户偏好，有效地提高搜索引擎查询的准确率。

2　用户偏好获取

2.1 用户信息收集与分析

QOMUR 采用显式和隐式相结合的方法来获取用户对

2006-05-31 收到，2006-11-13 改回
国家自然科学基金重点项目(60432010)和国家 973 计划项目(2003CB314806)资助课题

网页的偏好。用户在浏览网页时，可以显式给出对网页的评价，范围从 0 到 5，数值越大，表明用户对网页的兴趣程度越高。QOMUR 同时收集用户对网页的操作，隐式获取用户对网页的偏好：⑴浏览时间；⑵页面激活时，鼠标的移动时间；⑶鼠标在页面的点击次数；⑷拖动滚动条时间；⑸标记行为，包括添加到收藏夹、邮寄、打印、保存页面、复制。第(1)至第(4)类行为对用户偏好的影响用 4 个整型参数 $X_1\cdots X_4$ 进行描述,每类参数都有一个最大值，以避免单个参数对结果产生过大的影响。第⑸类行为对用户偏好的影响用 5 个布尔型参数 $X_5\cdots X_9$ 进行描述，当参数对应的行为发生时，参数取 1，否则取 0。

用户打开到关闭一个网页的过程称为一次隐式评价，用户在一次隐式评价中表现出的偏好表示为上面 9 个参数的函数：$Y=f(X_1\cdots X_9)$。QOMUR 用线性回归模型[7]分析 Y 和 $X_1\cdots X_9$ 的关系，得出 Y 的完整表达式。

2.2 基本用户偏好

本体[8]是共享概念模型的形式化规范说明，可以理解和表达为一组概念的定义及其相互关系。本体概念分为类、实例、属性：类描述领域内对象的类别，实例描述具体的对象，属性描述各种概念之间的相互关系。本体提供了描述某个领域的术语和概念，而本体知识库则使用这些知识来表达该领域的事实。QOMUR 基于关键字词频向量比较[5]，进行网页归类：对知识库中的实例 c_j，设生成其关键字词频向量的网页集合为 D_j，取 c_j 与 D_j 中网页的相关度最小值为 c_j 的相关度门限，记为 Min_j。当网页 d 与实例 c_j 相关度大于 Min_j 时，称网页 d 与实例 c_j 相关，反之无关。

通过网页归类，QOMUR 将用户对网页的偏好转化为对实例的偏好。由网页归类直接得到的用户对本体实例的偏好，称为基本用户偏好。设一次显式或隐式评价中，用户对网页的偏好为 p，与该网页相关的实例集为 η，用向量 $\boldsymbol{P}=(\mathrm{date},p_1,p_2,\cdots,p_n)$ 表示该评价对应的基本用户偏好，n 为本体知识库中实例的个数，p_i 为评价中实例 $c_i(1\le i\le n)$ 的偏好值，当 $c_i\in\eta$ 时，$p_i=p$，当 $c_i\notin\eta$ 时，$p_i=0$，date 为评价发生的时间。

3 用户偏好分析

3.1 语义关联和扩展用户偏好

在本体知识库中，如果两个概念之间存在一条或多条属性序列，称这两个概念存在语义关联[9]。语义关联的强度表明概念之间语义关系的紧密程度。QOMUR 依据继承关联和路径关联，分析用户对知识库中实例隐含的兴趣。

定义 1 (扩展用户偏好) 从基本用户偏好出发，通过实例间关联分析，得到的用户对本体实例隐含的兴趣。

本文假设对某概念具有偏好的用户，很可能对与此概念存在较强语义关联的其他概念也有偏好，借此优化查询的过程。在图 1 中，张明(概念 B)是论文网络的演进(概念 C)的作者，概念 B，C 之间存在较强的语义关联。用户阅读介绍论文网络的演进(概念 C)的网页 F 后，输入关键字“张明”查询，搜索引擎返回网页 D 和 E(分别关联到概念 A，B)，由 B，C 的相关性，网页 E 是用户查询的目标的可能性显然更大。这个例子中，用户对概念 B 的偏好即为扩展用户偏好。

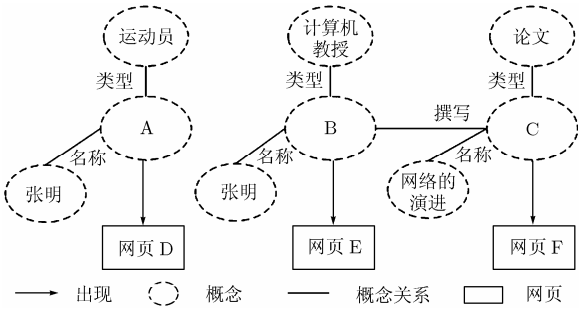

图 1 概念关系结构

3.2 用户偏好扩散

定义2(偏好扩散) 根据基本用户偏好，通过语义关联分析，发现扩展用户偏好的过程，称为偏好扩散。

定义3(局部深度最大的共同祖先) 类 C 是概念 a，b 的共同祖先，若 C 的子类中不存在 a，b 的共同祖先，称 C 是 a，b 的局部深度最大的共同祖先。

定义4 (重复的路径序列) 对属性序列 $\mathrm{ps}_1=\{a_1,P_{11},a_2,P_{12},a_3,\cdots,a_n,P_{1n},a_{n+1}\}$，$\mathrm{ps}_2=\{b_1,P_{21},b_2,P_{22},b_3,\cdots,b_n,P_{2n},b_{n+1}\}$，若 $\exists\, 1\le i,j\le n$，满足 $(\mathrm{a}_i=\mathrm{b}_j)\wedge(\mathrm{a}_{i+1}=\mathrm{b}_{j+1})\wedge(\mathrm{P}_{1i}=\mathrm{P}_{2j})$，称 ps_1，ps_2 为重复的路径序列，反之称 ps_1，ps_2 为不重复的路径序列。

设 $\boldsymbol{P}=(\mathrm{date},p_1,p_2,\cdots,p_n)$ 为基本用户偏好向量，n 为本体知识库中实例的个数，$p_i(1\le i\le n)$ 为实例 c_i 的显式偏好。$p_{ij}^{\mathrm{inherit}}$ 和 p_{ij}^{path} 分别为 c_i 通过继承关联和路径关联扩散到实例 $c_j(1\le j\le n,\ i\ne j)$ 的偏好，初始值为 0。从 $\boldsymbol{P}$ 开始的偏好扩散算法如下：

⑴从 c_i 开始继承关联扩散 设 c_i 信息量大于 ε 的祖先集合为 $A_i=\{C_{i1},C_{i2},\cdots,C_{im}\}$(类 C 的信息量 $I=-\log\Pr[C]$，$\Pr[C]$ 表示知识库中任一实例属于类 C 的概率)。对 A_i 中的每个元素 $C_{ik}(1\le k\le m)$，设 E_{ik} 是和 c_i 以 C_{ik} 为局部深度最大的共同祖先的实例集合，对 E_{ik} 中的所有实例 c_j，更新 $p_{ij}^{\mathrm{inherit}}=p_{ij}^{\mathrm{inherit}}+\alpha p_i\times\dfrac{H_{C_{ik}}}{H_{\mathrm{depth}_{ik}}}$。$\alpha$ 为继承关联扩散的衰减因子，$H_{C_{ik}}$ 为 C_{ik} 在本体层次中的深度，$H_{\mathrm{depth}_{ik}}$ 为 C_{ik} 所在分支的最大深度，$\dfrac{H_{C_{ik}}}{H_{\mathrm{depth}_{ik}}}$ 为 c_i，c_j 由于继承 C_{ik} 产生的语义关联的强度，C_{ik} 的层次越深，代表的意义越明确，实例 c_i，c_j 间的继承关联就越大，由 c_i 继承关联扩散到实例 c_j 的偏好也越大。

⑵从 c_i 开始路径关联扩散　从 c_i 开始广度优先遍历，记录经过的路径，保证各遍历分支的路径序列不重复，以避免出现循环。设某条遍历分支到达实例 $c_j(1\le j\le n,\ i\ne j)$时，路径长度为 step，更新 $p_{ij}^{\text{path}}=p_{ij}^{\text{path}}+\beta*\frac{p_i}{\text{step}^2}$，$1/\text{step}^2$ 为 c_i、c_j 由这条路径产生语义关联的强度，β 为路径关联扩散的衰减因子。c_i，c_j 间不重复的路径序列条数越多、长度越短，c_i，c_j 间的路径关联就越大，通过路径关联从实例 c_i 扩散到实例 c_j 的偏好也越大。为了保证程序的效率，规定遍历的最大深度 maxStep=3。

(3)综合基本和扩展偏好　重复(1)，(2)，对每个实例作偏好扩散。

扩散后的用户偏好记录为 $\boldsymbol{P}'=(\text{date},\ p_1',p_2',\cdots,p_n')$，其中 $p_i'=p_i+\sum_{m=1}^{n}(p_{mi}^{\text{inherit}}+p_{mi}^{\text{path}})$。下文中的用户偏好记录，如无特别说明均指扩散后的用户偏好。

3.3 用户偏好合并

QOMUR 根据用户在查询 T 之前一段时间内的偏好记录，生成综合用户偏好向量 $\boldsymbol{P}=(p_1,p_2,\cdots,p_n)$，以预测用户在查询 T 发生时的兴趣状态，其中 n 为本体知识库中的实例数，p_i 为用户对实例 c_i 的偏好的预测值。

按照用户偏好记录对应评价的发生时间，用户偏好分为历史用户偏好和会话用户偏好。会话是若干次连续的查询，由用户在查询过程中指明，会话内所有查询的对象相同或相近。历史用户偏好是用户偏好合并前 N 天的用户偏好记录(不包括查询 T 所处会话)，N 由用户指定。会话用户偏好是查询 T 所在会话的用户偏好记录。

取查询 T 发生的时间为第 0 天，前一天为第 1 天，依此类推，第 i 天的历史用户偏好为 $\boldsymbol{P}_i^{\text{his}}=(p_{i1}^{\text{his}},p_{i2}^{\text{his}},\cdots,p_{in}^{\text{his}})$，

$$p_{ij}^{\text{his}}=\frac{1}{S_i}\sum_{\text{hp}=1}^{S_i}p_j^{\text{hp}}\bullet e^{-\frac{\log 2}{\text{hl}}i},\quad 1\le j\le n \tag{1}$$

其中 S_i 为第 i 天的用户偏好记录总数；p_j^{hp} 是第 i 天第 hp 条记录中第 j 个概念的偏好值；hl 为偏好记录对综合偏好的影响力的衰减周期，如取 hl=7，则经过 7 天，该偏好记录的影响将衰减为初始值的一半。

N 天内的历史用户偏好 $\boldsymbol{P}^{\text{history}}=(p_1^{\text{history}},p_2^{\text{history}},\cdots,p_n^{\text{history}})$，

$$p_j^{\text{history}}=\sum_{i=1}^{N}p_{ij}^{\text{his}},\ 1\le j\le n \tag{2}$$

会话用户偏好 $\boldsymbol{P}^{\text{session}}=(p_1^{\text{ses}},p_2^{\text{ses}},\cdots,p_n^{\text{ses}})$，

$$p_i^{\text{ses}}=\frac{1}{S}\sum_{hp=1}^{S}p_i^{\text{hp}},\ 1\le i\le n \tag{3}$$

S 为当前会话用户偏好向量总数，p_i^{hp} 是该会话第 hp 个偏好向量中第 i 个概念的偏好值。

合并历史用户偏好和会话用户偏好，综合用户偏好向量为

$$\boldsymbol{P}=\alpha\times\boldsymbol{P}^{\text{history}}+\beta\times\boldsymbol{P}^{\text{session}} \tag{4}$$

其中 $0\le\alpha$，$\beta\le 1$，$\alpha+\beta=1$。由于即使在一个很短的时间段内发生的多次查询，对应的概念也可能不同，当前会话中的用户偏好能够最有效地反映出用户的真实目的，因而设 p_i^{ses}，p_i^{history} 分别为 $\boldsymbol{P}^{\text{session}}$ 和 $\boldsymbol{P}^{\text{history}}$ 中的最大值，α，β 的取值应满足 $\alpha\times p_i^{\text{history}}<\beta\times p_i^{\text{ses}}$。

4　查询优化方法

QOMUR 依据用户偏好将搜索引擎返回的网页重新排序。设网页 Page_i 为搜索引擎返回的第 i 个结果，M 为返回结果总数，记 Page_i 与用户查询请求的关键字相似度为 $M-i$。根据 Page_i 的标题和摘要，将其归类到实例集合 $C_i=(c_{i1},c_{i2},\cdots,c_{in})$。用户对 Page_i 的偏好 $W_i=\sum_{j=1}^{n}p_{ij}$，p_{ij} 为当前的综合用户偏好向量 $\boldsymbol{P}$ 中实例 c_{ij} 对应的元素值，即 W_i 为用户对与 Page_i 相关实例的偏好之和。综合用户偏好和关键字相似度，网页 Page_i 的兴趣度预测值为 $I_i=W_i+\gamma\cdot(M-i)$，当 $W_i\ne 0$ 时，$\gamma=0$，反之，$\gamma=1$。

将网页按兴趣度预测值大小降序排列，即为 QOMUR 查询优化的结果。

5　实验及分析

5.1 原型系统

为验证 QOMUR 查询优化的性能，本文实现了相应的原型系统 SASPWS。系统为图 2 虚线内的部分：用户通过用户界面提交查询请求，察看查询结果；用户信息采集模块监控用户对网页的各种操作，提供用户对网页进行显式偏好评价的接口；用户行为分析模块分析用户对特定网页的兴趣度；本体知识库是 SASPWS 推理的基础，SASPWS 采用了斯坦福大学的 TAP[10]本体库；网页归类模块计算网页与本体知识库中实例的相似度，进行网页归类(由于知识库中的实例数目过于庞大，目前网页只能归类到本体知识库中部分核心的实例)；语义关系分析模块根据概念之间的语义关联进行推理，发现隐含的用户偏好；用户偏好管理模块合并不同类别和时期的用户偏好，得到查询优化依据的综合用户偏好；个性化结果生成模块根据综合用户偏好，对网页重新排序。系统对传统搜索引擎的访问通过封装 Google Web APIs 实现。

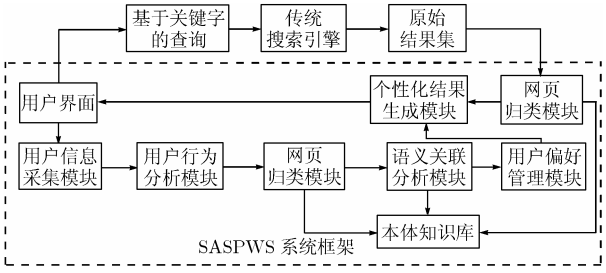

图 2　SASPWS 系统框架

下面通过与 Google 及个性化搜索方法 Micro[6](依据 ODP [11]顶层概念分类)的比较，验证 QOMUR 的有效性。在北京邮电大学计算机系征集 5 个志愿者，志愿者根据个人兴趣从 TAP 知识库选择本体实例，自拟关键字查询。连续收集 30 天的用户数据，将前 29 天的数据作为训练集，第 30 天的数据作为测试集。SASPWS 系统和 Micro 方法分别根据训练集，建立实验者的用户偏好模型。

5.2 查询优化效果比较

实验时，从每个实验者的测试集中随机抽取 5 个查询问题，通过 SASPWS，Micro 和 Google 查询，由问题的提出者判断结果的相关性。SASPWS，Micro 对每个问题查询两次(SASPWS 的两次查询在同一会话内)，实验者从第 1 次查询返回网页中选择浏览 3 个，再进行第 2 次查询。图 3 为前 20 个查询结果的准确率，Micro 的两次查询曲线重合，在图中合并为 1 条。几种方法的查询准确率由高到低依次为：$SASPWS_2$(SASPWS 第 2 次查询)，$SASPWS_1$(SASPWS 第 1 次查询)，Micro，Google。采用个性化技术的 SASPWS，Micro 方法，均起到了优化查询结果的作用，SASPWS 的优化效果优于 Micro，表明 QOMUR 比以往的用户偏好推理方法，能够更加准确全面地获取用户偏好。$SASPWS_2$ 较 $SASPWS_1$ 具有更高的查询准确率，表明当前会话的用户偏好，能更准确地说明用户查询目标。

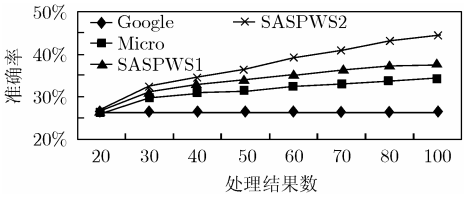

图 3 查询准确率 1

5.3 用户兴趣与查询效果

每个实验者从 TAP 知识库中选取 4 个实例，自拟关键字，查询实例的相关信息。查询过程、正确性判断方法与 5.2 节实验相同。图 4 为选出实例与该实验者在偏好训练阶段表现出的兴趣相关时，前 20 个查询结果的准确率(Micro 两次查询所得曲线重合，在图中合并为 1 条)，由高到低依次为：$SASPWS_2$，$SASPWS_1$，Micro，Google；图 5 为选出实例与该实验者在偏好训练阶段表现出的兴趣无关时，前 20 个查询结果的准确率(Micro 两次查询所得曲线重合，在图中合并为 1 条)。查询准确率由高到低依次为：$SASPWS_2$，Google，Micro，$SASPWS_1$。$SASPWS_2$ 相对于 SASPWS1 查询的准确率显著提高，分析前 80 个结果，SASPWS1 准确率 21.5%，$SASPWS_2$ 达到 43%，Google，Micro 相应的准确率为 28.3%、25.2%。

实验结果表明，当查询用户熟悉领域内的概念时，SASPWS 和 Micro 均能起到优化查询结果的作用，SASPWS 优化效果更为明显。当用户查询陌生概念时，由于概念间的干扰，$SASPWS_1$ 和 Micro 的准确率低于 Google。$SASPWS_2$ 查询准确率显著提高，表明 SASPW 能够迅速的发现并适应用户兴趣的变化，Micro 对用户兴趣变化需要较长的适应过程。

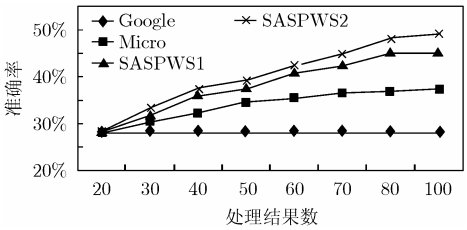

图 4 查询准确率 2

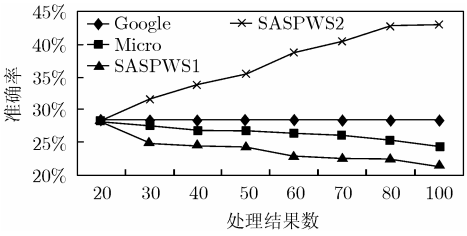

图 5 查询准确率 3

6 结束语

个性化搜索技术可以为不同兴趣和背景的用户提供满足其个性化需求的搜索服务。本文对在个性化搜索系统中如何有效地发现用户偏好进行了研究，提出了一种基于用户偏好分析的查询优化方法。该方法通过分析本体实例之间的语义关联，发现隐含的用户偏好；综合不同时期的用户偏好，建立用户行为和用户偏好之间的关系模型，能够快速地发现并适应用户兴趣的变化。实现了相应的原型系统，实验表明，相对于传统的个性化搜索方法，本文提出的方法可以有效地发现用户偏好，提高查询的准确率。在进一步的工作中，计划对用户偏好的扩散算法进行改进，考虑更多的信息，并对分类的技术进行优化和改进。

参 考 文 献

[1] Keenoy K and Levene M. Personalisation of Web search. IJCAI 2003 Workshop on Intelligent Techniques for Web Personalization, Acapulco, Mexico, 2003: 201–228.

[2] Konstan J, Miller B, and Maltz D, *et al.*. GroupLens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 1997, 40 (3): 77–87.

[3] Sugiyama K. Adaptive Web search based on user profile constructed without any effort from users. WWW2004, New York, 2004: 675–684.

[4] 蒋宗礼, 肖华, 赵钦. WebSifter: 个性化网络搜索辅助系统. 清华大学学报(自然科学版), 2005, 45(增刊): 1903–1907.

Jiang Zong-Li, Xiao Hua, and Zhao Qin. WebSifter: An assistant system for personal web search. *Journal of Tsinghua University* (*Science and Technology*), 2005, 45 (Sup): 1903–1907.

[5] Susan G and Jason C. Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems*, 2003, 1 (3-4): 219–234.

[6] Mirco S and Susan G. Personalized search based on user search histories. Proceeding of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, Compiègne, France, 2005: 622–628.

[7] 葛余博.概率论与数理统计. 北京: 清华大学出版社, 2005: 324–326.

[8] Gruber T R. A translation approach to portable ontology specifications. *Knowledge Acquisition* , 1993, 5(2): 199–220.

[9] Anyanwu K and Sheth A. ρ-Queries:enabling querying for semantic associations on the semantic web. Proceeding of the WWW2003, New York, 2003: 690–699.

[10] Guha R and McCool R. Tap: A semantic Web test-bed. *Journal of Web Semantics*, 2003, 1(1): 81–87.

[11] The Open Directory Project(ODP). http://dmoz.org.

梅　翔：　男，1979 年生，博士生，研究方向为语义网、信息检索.

孟祥武：　男，1966 年生，教授，硕士生导师，主要研究领域为网络软件、智能移动业务平台.

陈俊亮：　男，1933 年生，中国科学院和中国工程研究院院士，博士生导师，主要研究领域为智能网、交换技术、通信软件、下一代网络.