

# 一种基于用户兴趣的搜索引擎输入信息处理方法

丁 伟, 谢彦峰 张忠林

(兰州交通大学电子与信息工程学院, 甘肃 兰州 730070)

**摘 要** 搜索引擎是互联网普及的标志, 目前搜索引擎在查全率和准确率上是不能让用户满意的, 如何让用户获得有用的信息已成为信息检索系统急需解决的问题。文章主要对用户个性化搜索引擎系统进行了研究, 提出了一种基于用户兴趣的搜索引擎信息处理方法。首先对系统各模块的功能进行了详细的研究和介绍, 然后具体介绍了输入信息处理方法流程及相关技术的研究。实验结果表明, 通过对输入信息进行了关键词处理后, 可以有效地提高对用户所需信息的理解的准确率, 从而提高搜索引擎的查全率和准确率, 具有一定的理论价值和应用价值。

**关键词** 搜索引擎; 个性化; 输入方式; 反馈技术

**中图分类号** :TP391.3 **文献标识码** :A

## A New Importation of Information Processing Methods Based on User Interested

DING Wei, XIE Yan-feng, ZHANG Zhong-lin

(Lanzhou Jiaotong University Electronics and Information Engineering Institute, Lanzhou Gansu, 730070)

**Abstract** :The search engine is the Internet popular symbol, at present the search engine cannot let the user being satisfied in the recall and the rate of accuracy, the question how to cause the user to obtain the useful information becomes urgently for the information retrieval system. This article mainly conducts the research of the user personalization search engine system, proposed one kind search engine information processing method based on the user interest. Firstly, the article has conducted the detailed research of the system various modules' function and introduced then introduces the inputted information processing method flow and the correlation technique research specifically. The experimental result indicated that through carrying on key word processing after the inputted information, may be effective enhancement to the user need the information and the understanding rate of accuracy, thus raises search engine's recall and the rate of accuracy, has certain theory value and the application value.

**Key words** :search engine; personalization; input mode; feedback technology

## 0 引言

随着网民使用互联网熟练程度的不断增加以及互联网技术的不断发展, 搜索引擎技术两个主要的发展方向是实现个性化和智能化搜索<sup>[1]</sup>。个性化搜索主要是通过跟踪分析用户的搜索行为, 充分地利用这些信息来提高用户的搜索效率; 智能化搜索主要体现在以下两方面, 一是对搜索需求信息的理解, 二是系统具有自适应、自调节的能力。

目前的搜索引擎主要是采用以关键字输入为基础的检索<sup>[2,3]</sup>, 用户输入检索关键字向搜索引擎提出查询请求, 引擎根据关键字对网页索引库进行检索, 对检索结果按照一算法排序后返回。这种方式具有一定的局限性:

- (1) 用户难以清楚的表达实际需求信息;
- (2) 搜索引擎不能较好的理解自然语言输入信息。

实际上, 输入方式往往决定了整个搜索引擎系统的索引方式、检索方式, 甚至体系结构。抛弃自然语言理解技术不发达、用户对搜索引擎不熟悉等客观因素, 依然可以在挖掘用户认知能力、设置个性化环节等方面改进输入方式, 从而提高检索质量。本文通过对搜索引擎系统的改进, 提出了一种新的输入信息处理方法, 可以有效地提高搜索结果的查全率和有效率。

## 1 系统模块及其功能

本文的个性化搜索引擎系统主要包括用户代理模块、查询扩展和分析模块、独立搜索引擎接口模块、信息过滤模块、结果反馈模块、数据库模块等部分。

收稿日期 2008-06-18

作者简介: 丁 伟(1984-)男, 山东枣庄人, 硕士研究生, 研究方向为人工智能及信息检索; 谢彦峰(1951-)男, 高级工程师, 主要从事计算机教学与单片机方面的研究; 张忠林(1965-)男, 教授, 主要从事人工智能、数据挖掘与计算机网络方面的研究。

各个模块所要实现的功能如下：

(1)用户代理模块 主要是向系统发出请求和接受系统的查询结果,给用户提供一个友好的交互界面。根据用户的输入检索信息,结合搜索引擎的检索器的接受输入格式,对复杂的搜索引擎的语法进行研究并作相应的转换,从而使独立搜索引擎可以更好的理解用户的需求信息,更好的发挥独立搜索引擎的检索优势,这样就提高了搜索引擎的关键词语法转化能力,从而减少因语法转化造成的信息丢失。

(2)查询扩展模块 主要是根据用户兴趣库内容和信息反馈模块来对输入信息进一步进行归纳和综合整理,从而可以全面的理解和识别用户的实际需求信息,可以进一步提高搜索引擎的查全率和准确率。

(3)独立搜索引擎接口模块 根据用户查询的信息内容不同以及各个搜索引擎的查询优势不同,合理的选择独立搜索引擎进行搜索查询。

(4)信息过滤模块 实现信息过滤,根据信息过滤算法和用户兴趣库对独立搜索引擎返回的信息检索结果做进一步处理,去掉重复文档并按相关度排序后提交给用户,过滤掉与用户背景知识无关或用户不感兴趣的信息,提高查准率。

(5)结果反馈模块 主要是根据用户对查询结果的查看以及评价信息,对查询结果进行分析和归纳,并把分析结果作出相应的处理,将用户输入感兴趣的内容信息传递到用户兴趣库中,并把语义相关的关键词保存到关键词库中,根据相应的算法作相关度处理。搜索引擎可据此信息过滤下次搜索的不相关或相关度不大的检索结果,精简检索结果,并且从用户反馈和检索结果中提取用户偏好信息,动态地修改用户兴趣库和语义相关库。

(6)数据库模块 主要包括两部分：

(1)用户兴趣库 为了提供面向用户的检索,系统必须维护用户的相关特征。一般地,用户信息应该包括用户感兴趣的主题、浏览方式、用户的领域知识、用户目标、背景、使用经验以及相应爱好等。

(2)语义相关库 建立语义相关信息库,主要是通过各种反馈技术与数据挖掘技术相结合来对同一关键词作进一步分析,从而得到更多的相关语义信息,进而准确理解用户的需求信息,提高搜索引擎的查全率。

## 2 系统结构图及信息处理流程

本文所研究的个性化智能搜索引擎系统结构图如图 1 所示。

其工作流程如下：

- (1)用户输入查询信息。
- (2)根据输入信息进行关键词拆分,调整查询表达式。
- (3)结合用户兴趣库和语义相关库更改关键词内容,调整用户查询表达式。
- (4)通过检索器(独立搜索引擎)来搜索相关信息。
- (5)依照用户兴趣来分析过滤检索结果。
- (6)将检索结果返回给用户。
- (7)根据用户的查询行为反馈去更新用户兴趣库及相关语义库。

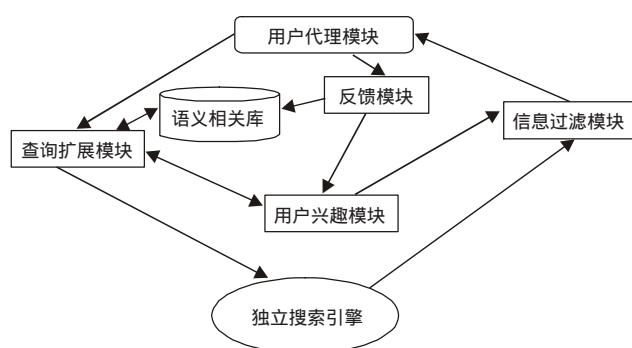


图 1 系统结构图

其中输入信息处理部分的流程图如图 2 所示：

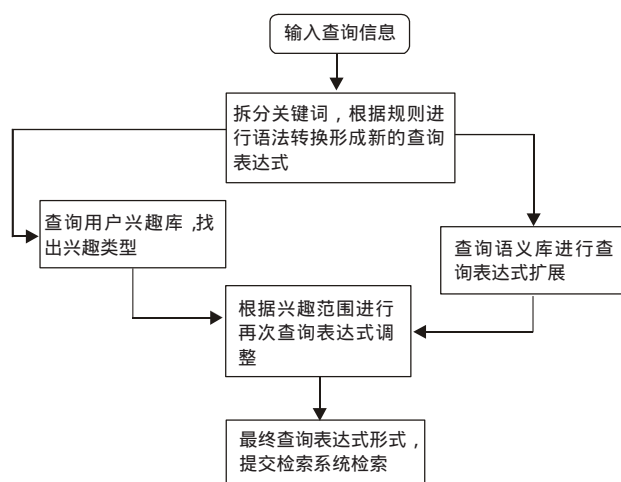


图 2 关键词处理流程图

## 3 应用技术的研究及实现

### 3.1 用户兴趣模型的建立

此部分的主要目的是建立用户兴趣库<sup>[4]</sup>,主要又可分为以下几部分：

- (1)兴趣匹配模块。
- (2)兴趣初始化模块(包括角色兴趣初始化模块和

用户兴趣初始化模块)。

(3)兴趣管理模块(包括角色兴趣管理模块和用户兴趣管理模块)。

(4)兴趣优化模块(包括角色兴趣优化模块和用户兴趣优化模块)。

用户兴趣模型图如图 3 所示：

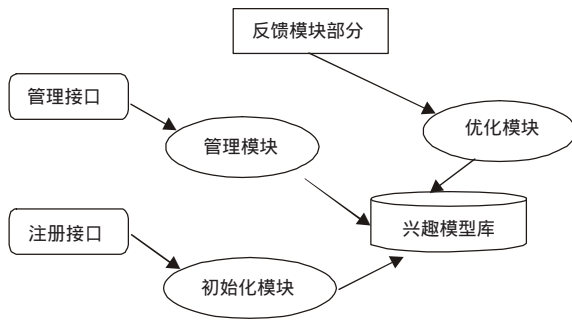


图 3 用户兴趣模块结构图

### 3.1.1 兴趣初始化模块

#### 3.1.1.1 角色兴趣初始化

在系统初始化的时候,我们通过给每个角色提供一系列的感兴趣文档  $RD=\{D_1, D_2, \dots, D_n\}$ , 然后对该角色的兴趣模型进行初始化。初始化的过程就是根据这些感兴趣文档, 采用下面的算法产生角色的兴趣集  $RV=\{V_1, V_2, \dots, V_m\}$ 。

```

RV=Φ // 兴趣集置空
For Di in RD// 根据反馈信息对用户兴趣进行权值和内容的更新
  Flag=FALSE
  For Vj in RV
    If P(Di, Vj) > λ
      Vj = Vj + P(Di, Vj) * Di
      Flag=TRUE
  if flag==FALSE
    Rv = Rv ∪ {Di}

```

在上面的算法中,对于每个感兴趣文档  $D_i$ ,我们采用 VSM 的相似度计算感兴趣文档  $D_i$  和角色的每个兴趣  $V_j$  的匹配度,如果匹配度大于阈值  $\lambda$ , 则将文档  $V_j$  的词向量(乘以一个权重  $P(D_i, V_j)$ )增加到兴趣  $V_j$  的向量中。如果感兴趣文档和角色的所有兴趣的匹配度都低于阈值, 则从该感兴趣文档产生一个新的用户兴趣, 并将该用户兴趣增加到用户兴趣集  $RV$  中。

在这个算法中,对阈值  $\lambda$  的设置尤为关键。阈值设置得太小,则产生的兴趣太少。阈值设置得过大,则产生的兴趣过多。可以看到,随着阈值的减小,用户的每个兴趣覆盖的兴趣文档也越多,更能反映用户的兴趣,因此能有效地提高查询的查全率。但同时,由于覆盖的兴趣文档越多,引入的噪声也越大,将会导致查准率降低。只有选择恰当的阈值,才能够做到有效地对用

户的兴趣进行表示和基于用户兴趣进行有效的查询。

#### 3.1.1.2 用户兴趣初始化

在系统中,用户兴趣初始化通过用户注册服务来实现。当用户注册系统时,需要设置自己所属的角色的类型,以便初始化该用户的兴趣。用户的初始兴趣集是用户所设置的所有角色的兴趣集的并集。设用户所设置的角色为  $R_1, R_2, \dots, R_n$ , 则用户初始兴趣集为  $UV=R_1V \cup R_2V \cup \dots \cup R_nV$ 。

#### 3.1.2 用户兴趣管理模块

提供了用户兴趣管理模块,用户可以查看自己当前的兴趣,以及每个兴趣词向量信息,可以手工地增加、修改和删除某个兴趣的词向量或者某个兴趣。通过这种途径,高级用户能够有效地调整自己的兴趣模型,从而使自己的兴趣模型尽量地符合自己的需求。同样设置一个角色兴趣管理模块,管理员可以手工地调整每个角色的兴趣及兴趣的词向量,从而提高角色兴趣模型的准确性。

#### 3.1.3 用户兴趣优化模块

在系统中,通过运行一个后台程序,定期扫描用户/角色兴趣集中的兴趣,将那些相似的兴趣(相似度大于某个阈值)合并为一个大的兴趣,从而减少用户兴趣的数量,提高用户兴趣的质量,优化算法如下:

```

For Vi in UV
  For Vj in UV
    If P(Vi, Vj) > γ
      UV = UV - {Vi, Vj} + {Vi+Vj}

```

### 3.2 相关反馈技术的研究

兴趣相似的用户也会有不同的侧重点,并且用户的兴趣随着时间会渐进变化,因此使用静态的用户兴趣模型往往不能完全反映用户的兴趣。反馈模块的作用就是让用户通过自己的检索行为或其他方法对自己的兴趣模型和相关语义进行有效的修正,使自己的用户兴趣模型越来越接近自己的实际兴趣,并且使系统语义库信息更加完善。系统获得反馈的途径可以有多种方式<sup>[9]</sup>, 本文主要对以下两种方法进行研究。

#### 3.2.1 点击反馈技术

用户对检索结果的点击是一种正向的反馈,表示用户认为这条检索结果切合他这次的查询意向,因此我们采用被点击的检索结果的向量更新用户的兴趣模型,其算法类似于角色兴趣模型的初始化算法,即如果被点击的检索结果  $D_i$  和用户的某个兴趣  $V_j$  比较相似,则将该检索结果的向量(乘以相似度作为权重)增加到兴趣  $V_j$  中;如果检索结果  $D_i$  和用户的所有兴趣都不相似,则采用  $D_i$  初始化一个新的兴趣  $V$ , 并加入到用户的兴趣集中。

对于用户所属的角色  $R$ , 我们采用类似的方法, 用  $D_i$  的向量更新角色  $R$  的兴趣模型。

同样类似利用网页的深层连接去挖掘相关信息, 从而进一步结合查询相关语义库信息去更新相关语义库信息。

### 3.2.2 评分反馈

用户在点击某条结果后, 可以对该结果进行评分, 用好、中、差来表示。我们采用和点击反馈类似的方法将文档  $D_i$  的向量更新用户的兴趣模型, 即将该被评分的文档  $D_i$  的向量, 乘以评分的值作为权重, 增加到用户的相似的兴趣  $V_j$  中。可以看到, 采用这种方法, 用户能够对由于点击而造成的错误反馈进行反悔, 并对点击造成的正确反馈进行鼓励, 从而进一步提高用户兴趣模型的准确性。如果用户不对点击结果进行评分反馈, 则我们认为用户的评分反馈是中。对于用户所属的角色  $R$ , 我们采用类似的方法, 用  $D_i$  的向量更新角色  $R$  的兴趣模型。

通过类似的评价方法去修改语义相关词的相关度, 根据每次点击信息去增加或减少相关度, 从而使系统更好地理解用户的实际需求。

## 4 实验设计

实验平台的搭建主要分为以下几部分:

(1) 独立搜索引擎的部分采用百度搜索引擎进行信息检索实验;

(2) 输入信息分词部分使用中国科学院计算技术研究所的中文分词系统 ICT-CLAS 对输入信息分词处理;

(3) 关键词处理扩展算法部分采用  $c\#$  语言 .net 开发功能模块;

(4) 用户兴趣库以及相关信息采用 xml 文档格式进行存储;

(5) 反馈技术部分采用点击反馈和评分反馈相结合进行试验;

实验数据部分主要分为:

(1) 取 5 种不同兴趣的用户进行实验;

(2) 对每一用户, 选取不同的输入信息内容进行输入检索实验, 内容如下:

输入信息查询内容	所要查询的意图
1 深蓝更蓝	查询计算机发展方面的信息
2 购买笔记本	查询笔记本厂商, 性能, 报价等方面的信息
3 白菜	查询蔬菜相关信息
4 Sql	查询数据库相关信息
5 安全教育	查询自我保护和能力培养方面信息
...	...

实验结果数据的比较分为:

(1) 同一检索内容, 不同用户角色检索结果的比较;

(2) 同一用户, 同一检索信息在百度上检索与在此实验系统运行下检索结果的数据比较;

在系统初始运行时, 各种数据比较差距不大, 随着系统的逐渐训练的运行一段时间后, 多次实验结果比较表明, 经过关键词处理后的检索结果较原始输入检索结果查准率和用户满意度有明显的提高。

## 5 总结

本文提出了一种基于用户兴趣和相关语义库信息, 并结合各种相关反馈技术的搜索引擎系统模型, 主要对个性化搜索引擎的输入信息处理上进行了改进。基于这种输入处理机制, 在一定程度上可以实现智能输入识别功能, 从而更加有效的提高关键词查询的准确率和查全率。

参考文献:

- [1] 陈金阳, 蒋建中, 郭军利, 等. 一种带反馈自适应的搜索引擎系统结构的研究[J]. 计算机与网络. 2003(23):54-58.
- [2] E.M.Voorhees, D.K.Harman, editors. Proceedings of 5th Text Retrieval Conference, Gaithersburg, Maryland, USA, November, 1996.
- [3] S.Wu, F.Crestani. Data Fusion with Estimated Weights. In Proceedings of the Eleventh International Conference on Information and Knowledge Management, McLean, USA, November 2001.
- [4] 肖卓程, 荆金华. 基于用户兴趣的搜索引擎[J]. 计算机应用与软件. 2007, 24(09):134-136.
- [5] 殷亚玲, 张蕾. 搜索引擎中语义相关反馈技术的研究[J]. 计算机技术与发展. 2006, 16(02):167-170.
- [6] 韩坤, 王剑锋, 崔忠强. Internet 的个性化搜索引擎的分析与研究[J]. 现代计算机. 2005, 10(221):31-34.
- [7] 朱素媛, 马溪俊, 梁昌勇. 人工智能技术在搜索引擎中的应用[J]. 合肥工业大学学报. 2003(26):657-661.
- [8] 徐宝文, 张卫峰. 搜索引擎与信息获取技术[M]. 北京:清华大学出版社, 2003.
- [9] Shen Xuehua, Tan Bin, Zhai Chengxiang. Implicit user modeling for personalized search [C]. Bremen, Germany: Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05), 2005. 824-831.
- [10] 印鉴, 陈亿群, 张钢. 搜索引擎技术研究与发展[J]. 计算机工程. 2005(14):54-56, 104.