基于隐私保护的个性化搜索系统

王金德 20721156

2008-01-20

目 录

第 1 章　背景和意义

互联网上信息飞快地增长，已经远远超过了人所能检索的范围。同时互联网上也充斥着大量冗余无用的信息，用户很难从这些无序的信息中提取有价值的信息。随之诞生的是日益强大的搜索技术。这些技术帮助用户搜索信息，解决了一部分信息爆炸所带来的问题。然而，不同的用户往往有不同的需求，不同的兴趣；同一个用户在不同的时候不同的场合也需要不同的信息，这些是通用的搜索技术所无法解决的。于是，一个新的研究和应用方向，个性化搜索发展起来[1]。

个性化搜索是提高网络搜索质量，增强用户体验的有效方法，并且已经成为搜索领域研究和应用的重要方向。但是要实现个性化网络搜索，往往需要获得用户的个性信息。而这些信息常常包含大量的用户隐私信息。在现代社会，越来越多的人开始关注隐私保护。伴随着各种先进技术的诞生，人生越来越开始担心个人隐私会不会泄漏。所以一个没有隐私保护的个性化搜索，其价值将大大降低。

不过隐私不是绝对的。不同的用户对隐私有不同的理解和定义[2]。而用户也常常愿意暴露少量不甚重要的隐私信息，而获得较好的个性化服务。隐私的不精确的特性给隐私保护带来了一定困难，同时也带来了很多机会。同时信息隐藏等诸多计算机安全方面技术的发展和成熟，为隐私保护提供了技术基础。

基于隐私保护的个性化搜索系统正是在隐私保护的基础上，设计的一个个性化搜索系统。它实际上是隐私保护和个性化搜索之间的权衡。

该搜索系统的意义在于，它与传统的搜索引擎不同，它把搜索技术和计算机安全方面的技术结合起来，在相对保护用户隐私的同时提供个性化搜索。同时，它对用户的隐私进行了量化的定义，使得用户的隐私保护程度变得可配置。论文将构造一个完整可行的系统设计方案，提出了一个具体的架构，并对架构中各部分的关键技术进行了分析和解决。

第 2 章　需求规格说明

2.1　目标和性能

简单地说，系统的目标是实现一个基于隐私保护的个性化搜索系统，提供良好的用户交互界面，保证较好的性能、安全和健壮性。

目标的细节如下：

- 系统在实现个性化搜索的基础上，重视隐私保护的实现。通过多种不同方式，不同层次的隐私保护机制，使得隐私保护的实现尽量少地影响个性化搜索的质量。
- 系统基于B/S架构实现，同时提供用户相应的浏览器插件，以进行更强功能的操作和更方便的配置。用户界面设计在友好简洁实用的基础上，进行一定的美化。
- 系统保证较好的性能。在功能设计和架构设计中就考虑性能，保证较高的个性化搜索性能，同时保证隐私保护也不会对性能带来太大影响。
- 保证较好的安全和健壮性。主要通过应用已有的像SSL之类的安全协议，以及设计健壮的架构和控制程序质量来保证。

性能是一个搜索系统相当重要的因素。本系统关注于隐私保护和个性化搜索，并不想独立地从基础开始去实现一个完整的搜索系统，它是在已经存在的搜索引擎的基础上实现的，比如利用Google API去访问Google的搜索服务，取得搜索结果。所以这部分的性能由被使用的搜索引擎决定。

本系统的性能主要体现在个性化信息收集、树状模型的建立、搜索页面的分类和二次重新排序。同时，对于个性化信息的网络传输也是影响系统性能的一个因素。而在这些方面，系统需从功能设计阶段就进行考虑权衡，把性能作为一个重要要素，使最后实现的系统性能尽量的高。

2.2　功能描述

系统的功能主要包括：

- 系统从用户个性化信息相关的多个数据来源，提取个性化信息，进行一定

的泛化和抽象，建立层次化的树型结构模型，把个性化信息按照其抽象程度组织起来，并由此对用户隐私进行量化。在树型结构模型建立之后，系统仍实时地自适应地根据最新的用户个性化信息对模型进行调整。

用户个性化数据的收集常分为隐式和显式两种[3]，本系统主要通过隐式方式收集，并提供少数几个可供用户显式配置的参数。隐私量化基于这样一种思考：越是抽象的东西其包含的隐私程度越小[4]。这样用户的隐私将被分为若干个级别，同时也提供更精确的数值化的隐私程度。用户可以根据自身的需要，选择或设置不同的隐私程度，以获得不同层次的个性化服务。

另外，数据来源包括用户访问历史记录[5]，用户收藏（包括网上文摘，如CSDN，delicious等），搜索时提供的关键词，以及对搜索结果的点击，并且这些数据来源是可配置的。

- 系统提供用户可配置的基于关键词过滤另一重隐私保护机制，保证一些敏感的词汇不会泄漏。关键词分为两种，一种是可以保存在客户端的树型结构模型中，但不会发送给服务器；而另一种在客户端也不保存。用户可以添加或删除关键词。

- 系统根据用户个性化信息，对一般的Web搜索结果（如Google返回的结果）进行重排序。这种排序首先在服务器端进行，然后再客户端进行第二次的重排序。因为为了保护隐私，客户端保存的用户个性化信息往往比发给服务器的要详细得多；而服务器进行的排序的页面却比客户端要多得多，因为由于网络、性能等原因，客户端不可能在收到所有服务器端页面之后再进行排序，而只能对一部分页面，比如20页进行更精确的重排。

- 提供用户一个良好的交互界面。用户通过Web浏览器进行搜索操作。并且以浏览器插件的形式，提供用户配置界面。

- 为了简化用户配置，系统对所有的配置信息，根据一般用户的普遍状况设置默认值。这种默认值不是固定的，系统会定期调整默认值，以更好地适应更多的用户。用户可以选择实时更新默认值，也可以选择固定不变。

第 3 章　技术方案

3.1　环境和技术

本系统将采用B/S架构，这样用户可以通过各种浏览器很容易地访问本系统提供的搜索服务。另外，由于用户个性化信息需要存储在客户端，并且需要提供用户配置各种参数的界面。为了较好的完成这一功能，浏览器的插件形式是一个很不错的选择。当然由于目前浏览器众多，而IE和FireFox浏览器最为广泛使用，并且它们都提供了较好地开发插件的环境，所以首先仅开发这两个浏览器的插件。

系统的基本架构和发布的形式，部分地决定了开发应用环境和所能选择的技术。从语言的选择来说，系统B/S架构的搜索服务部分使用J2EE（Java）开发，而浏览器插件部分则使用C++开发。另外，客户端需要进行用户个性化数据的采集、处理、建模等操作，还要进行页面的二次重排，浏览器客户端脚本（如JavaScript）和本地程序（比如ActiveX等）的联合开发也必不可少。

搜索系统服务部分将部署在Tomcat容器上。系统服务器端数据使用MySQL存储，而客户端的个性化数据以XML配置文件保存。

对于实际的网络搜索部分，系统利用Google API提供的搜索服务获取搜索结果，而不是自己去实现。

3.2　系统实现

3.2.1　架构

本小节主要讨论搜索系统主体部分的架构，并不讨论用户用来配置个性化信息和隐私相关数据的浏览器插件的结构，因为这些插件本身结构是很简单的。搜索系统架构如图3.1所示。

由图中可以看到，一个基于隐私的个性化保护主要分为如下几步：

1. 用户通过浏览器发出搜索请求。
2. 请求被客户端脚本（如JavaScript脚本）截获，它从客户端个性化信息文件

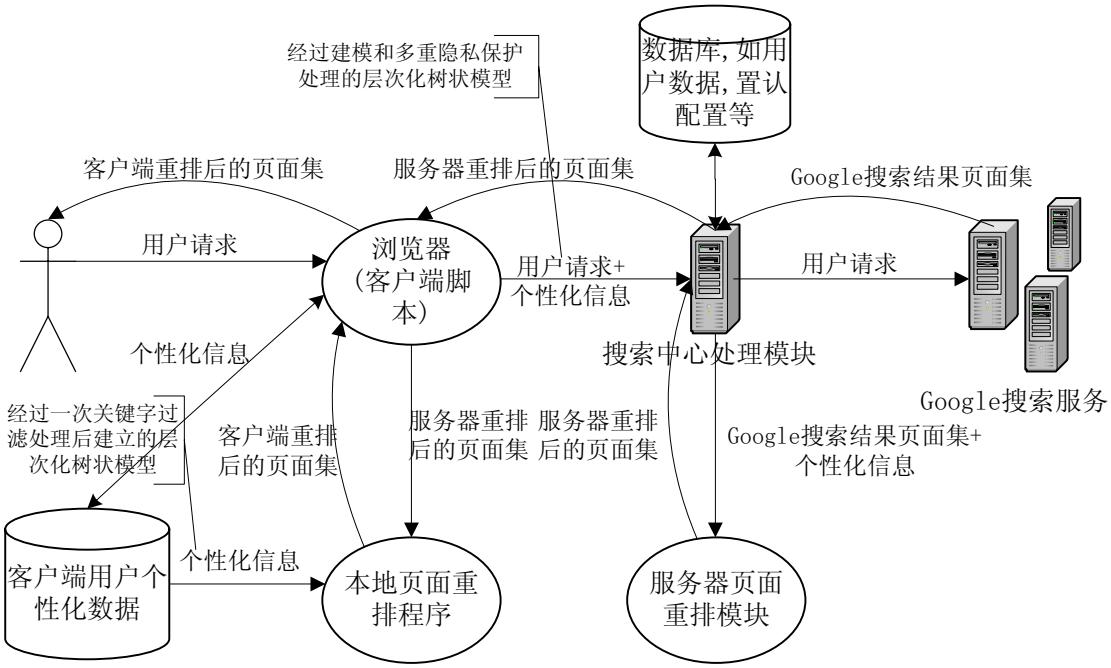

图 3.1 搜索系统架构

中获得需发送给服务端的个性化信息，其形式为经过建模和多重隐私保护处理的层次化树状模型。随后，脚本把用户请求和搜索请求一起发给服务器。

3. 服务器把用户请求发给Google搜索服务获得Google的搜索结果。
4. 服务器把返回的结果页面集与用户个性化信息，交给服务器页面重排模块进行权衡综合，进行页面重排序。
5. 服务器把排好的页面集发送给浏览器（客户端脚本）。
6. 客户端脚本取出页面集，交给本地的页面重排程序，用更详细的个性化信息对页面进行重排。这次的个性化信息是经过一次关键字过滤处理后建立的层次化树状模型，未经后序的一些隐私保护操作。
7. 重排好的页面最终显示给用户。

3.2.2 个性化信息建模过程

个性化信息建模的过程，极大地体现了本系统的隐私保护机制。该部分和页面服务器/客户端两次重排机制一起，构成了本系统隐私保护的核心内容。个

性化信息建模和隐私保护的过程如图3.2。

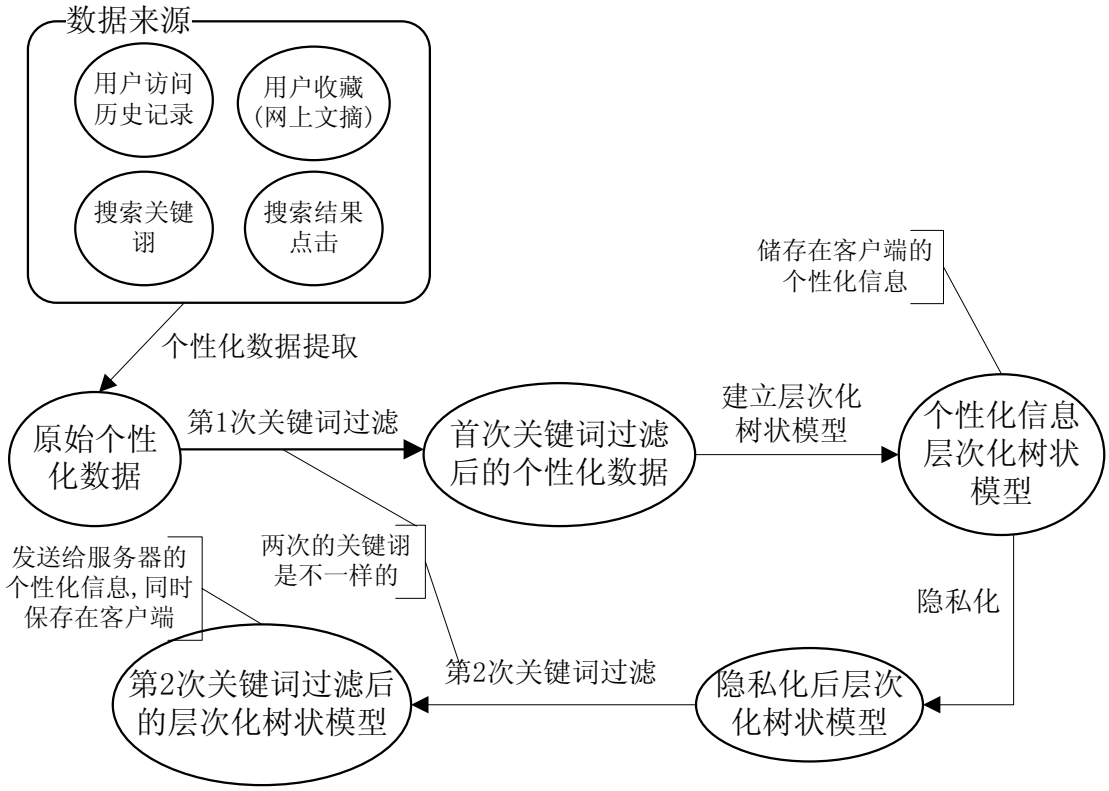

图 3.2 个性化信息建模和隐私保护处理

由图中看出，个性化信息建模及其中的隐私保护机制的实现主要分如下几步：

1. 根据用户配制，从数据来源（包括用户访问历史记录、用户收藏（包括网上文摘，如CSDN，delicious等）、搜索时提供的关键词、以及对搜索结果的点击）提取原始个性化数据。
2. 原始个性化数据经过第一次关键词过滤。
3. 个性化数据经过一定的清洁、抽象、泛化，根据抽象的程度建立层次化树状模型，并存储在客户端。
4. 层次化树状模型中所包含的隐私将被量化，再根据用户的配置对一些隐私进行隐藏、删除等。
5. 树状模型进行第二次关键词过滤，并且与前一次过滤的关键词是不同的。过滤后的个性化信息保存在客户端，将伴随用户搜索请求发给服务器。

第 4 章　关键技术分析和解决

实现本系统可采用的技术很多，不同的技术有着不同的特性，常常解决不同的问题，也有着各种不同的缺陷。本章将根据本系统的特点、性能，及各种技术本身的特性，进行权衡取舍，合理地实现各种功能。

同时，由于系统未曾实现，所以对于关键技术的分析和解决都是思想性的，常常会基于一些假设。当然，这种分析和解决在实际中并不一定成立，但能提供一个重要的参考和指导。并且，在实际开发和测试中，这些技术将被逐渐使用。

本系统使用的关键技术主要针对于两类功能：个性化信息建模和隐私保护（4.1），页面排序技术（4.2）。

4.1　个性化信息处理技术

本节的技术主要基于如下假设：

- 不同的兴趣在更高抽象层中可能是一样的[6]。
- 越概括的兴趣往往是越长久的兴趣，越具体的兴趣往往是越短期的兴趣。
- 根据用户行为的上下文在准确定位用户兴趣方面很重要，常可以消除自然语言的歧义。关于这一点，这里指的是当把一系列网页分类，实现概括兴趣和具体兴趣的层次结构。对于具体兴趣，当它位于概括兴趣的上下文之下时，其语义常常就不会有歧义。比如Apple在Computer的上下文之后，就会被当作苹果电脑，而不是苹果这种水果。

同时，应该注意的是，这些假设在一定条件是合乎情理的，但并非总是如此。

实现个性化信息处理技术主要包括个性化信息抽取、建模和隐私保护技术。

4.1.1　个性化信息抽取

对于不同的数据来源往往使用不同的抽取技术。

对于用户查询的关键词，基本不用做太多处理。因为它们很简短，并且常常本身就是关键词。

对于用户历史记录和收藏，系统先提取网页摘要信息，因为它们中常常包含非结构化的数据，并且直接对网页文本进行操作，常常因为文本模型的维度过高而性能低下。提取网页的技术一般包括[7]：

- Luhn's Summarization Method：是一种基于关键句子提取的方法。它对每个句子定义一个重要系数，重要系数高的句子被用来形成摘要。重要系数的计算分两步：先取词频在上下阈值之间的词建立重要词库；再根据词库计算每个句子的重要系数。
- Latent Semantic Analysis：其特点在于它用非常高维的“语义空间”中的点来表示词条和相关内容。
- Content Body Identification by Page Layout Analysis：根据余弦相似度，找到核心功能（句子），提取摘要。
- Supervised Summarization：这种方法先从网页中提取一系列特征，再应用一种有监督的机器学习算法，训练摘要提取器是否把某一句子加入到摘要中的鉴别能力。再利用摘要提取器进行提取。

这些方法各有特点，而论文[7]把四种方法综合在一起提高摘要质量。而本系统也将使用这种方法。

而对于用户对于搜索结果的点击选择，由于搜索结果本身是摘要信息，所以不需重新提取。

这样最原始的个性化信息就包括一些关键词和一些摘要信息。当然可以在些过程对这些个性化信息进行一定的清洁处理。

4.1.2　建模和隐私保护技术

建模和隐私保护技术主要包括三个技术：模型构建、隐私量化和关键词过滤。

模型构建 论文[4]描述了一种构建个性化信息层次化树状模型的方法。该方法基于这样一个假设：频繁出现的条目是用户感兴趣的话题。实际上，这种假设在很多系统实现中都存在，并且具有较高的合理性。

系统将借鉴这种方法来实现。不过条目提取的数据为4.1.1的结果：现成的关键词条目和摘要信息。这一步不采用原始网始数据的原因是网页异构化程度太高，并且直接使用网页效率无法保证。

对于关键词条目的提取，主要采用了分词技术和文本分类的一些相关技术[8]。

关键词条目提取之后，系统设置一个阈值从这些条目中选中频繁条目，这些频繁条目将被用来构造一个基于条目的个性化层次结构树状模型。模型至项向下生成，主要关注各个条目之间的联系，而不是简单把所有条目分组。条目之间的关系概括为相似条目、父子条目和不相关条目。关系的生成可以简单地通过统计词频和词在不同文档中的分布情况实现，方法与TF*IDF模型的思想类似。根据这些关系，树状模型可以很容易的被构造出来。

在构造树状模型的同时计算每个结点的支持度（Support），这主要用于量化隐私。

隐私量化 通过引进信息论中Self-Information或Surprisal的概念来衡量隐私程序[4]。根据概率论和信息论的结果，并且基于越具体敏感的条目，其Self-Information越大的假设，就可以使用每个条目（结点）的支持度来衡量隐私程度。这样，可以引进一个参数，表示用户希望公开其隐私程度的阈值。这个阈值与支持度存在一个函数关系，这样阈值的提供，可以灵活把一些隐私程序不高的树状部分提取出来。这实现也相当于一个截枝的过程，把过于敏感的关键词截掉。

同时，用户也可以较为明确地知道个性化隐私暴露的状况。

关键词过滤 需要进行关键词过滤的原因在于，虽然经过了隐私量化对隐私程度进行了设置，但是这只是去除了一些因为具体而敏感的条目。像另外一些非常抽象的条目，或者某一方向上的兴趣，用户可能仍然不想让人知道，最常见的比如Sex，Women之类。而这种条目是无法通过抽象泛化来过滤的。而系统或用户可以设定一些类似的关键词，这样，这方面的隐私也得到了保护。

3中已经提过系统将进行两次关键词过滤，并且两次过滤的关键词是不一样的。这是因为有些词用户可能仅仅不想发给服务器，而只是想存储在客户端以获得更好的个性化搜索服务。系统的第一次过滤，过滤掉用户不想

存储在客户端的数据，这部分数据也不会用来构建层次化树状模型。实际上，系统也可以用关键词过滤来排除一些干扰词。第二次过滤则过滤掉用户不想发给服务器端的词。

关键词过滤技术实际上是相当简单的，本身基本没有什么复杂的算法。可以直接遍历被过滤的数据，匹配关键词库并除去匹配的数据。

4.2　页面排序技术

系统进行了两次页面重排，分别在服务器和客户端进行。这两次重排所用的技术基本上是类似的，只是页面数量和个性化信息不同。

由于Google服务器返回的结果是已经排好序的，所以两次重排都是在原有排序基础上加入个性化信息进行调整。调整的算法很简单：

1. 取若干条搜索结果页面摘要，对这些摘要同样进行生成个性化信息层次化树状模型类似的操作，生成页面的树状模型。
2. 把结果与个性化信息层次化树状模型进行比较。比较的方法是应用余弦相似度公式比较两者的相似度，然后利用相似度生成一个初步的纯个性化排序。
3. 把纯个性化排序与输入的已排好序的页面进行线性加权合成，生成新的排序。

这种排序方法很简单，效率很高，不过要取得更好的效果，排序时的加权比重、选取的页面数目等都要在实验中调整。另外，也可以考虑一些更复杂的加权方式。

第 5 章 总结

随着互联网搜索技术的不断更新，用户对搜索质量要求的不断提高，个性化搜索成为一个热门的主题。个性化搜索技术一般包括用户描述文件的表达与更新、资源描述文件的表达、个性化推荐技术、个性化服务体系结构等[3]。本系统正是提供了这些技术的一种实现。

随着技术的日新月异，人们对隐私保护方面的要求也日益提高。一个无法保护用户隐私的个性化搜索系统是不完备的。本系统正是出于这一原因，把隐私保护作为系统最重要的原因之一加以考虑。

基于隐私保护的个性化搜索系统就是在一定程度隐私保护的基础上，实现个性化搜索，提高搜索质量，尽量得完成个性化搜索和隐私保护之间的均衡。

另外，本系统中涉及的各方面技术都在不断被研究改进。作为一个实际的系统实现设计，不可能对每一部分，找到最优方法，并且，也不一定有最优方法存在。所以该搜索系统采用了很多现成的技术，如文本分类、层次化树状模型建立、关键词过滤、页面排序等等，把这些技术综合在一起，获得最佳功能和性能。系统也没有完全自主独立地实现一个搜索系统的所有部分，而是直接利用一些现有的资源，比如Google提供的搜索服务等。

总体来说，该系统的优势在于个性化搜索和隐私保护的良好结合，在保护隐私的基础上尽可能地提供强大的个性化搜索功能。

参考文献

[1] Ferragina P and Gulli A. A personalized search engine based on web-snippet hierarchical clustering. Special interest tracks and posters of the 14th international conference on World Wide Web (WWW), 2005

[2] Xiao X and Tao Y. Personalized privacy preservation. Proceedings of the 2006 ACM SIGMOD international conference on Management of data (SIGMOD), 2006

[3] 曾春, 邢春晓. 个性化服务技术综述. 软件学报, 2002, 13

[4] Xu Y, Zhang B, Chen Z, et al. Privacy-enhancing personalized web search. Proceedings of the 16th international conference on World Wide Web, 2007

[5] Speretta M and Gauch S. Personalizing search based on user search histories. Conference’04, 2004

[6] Kim H R and Chan P K. Learning implicit user interest hierarchy for context in personalization. Proceedings of the 8th international conference on Intelligent user interfaces (IUI), 2003

[7] Shen D, Chen Z, Yang Q, et al. Web-page classification through summarization. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR), 2004

[8] 高洁. 文本分类技术研究. 计算机应用研究, 2004