基于自然语言理解的智能化多媒体信息检索系统研究*

师东生

(内蒙古科技大学信息工程学院，内蒙古 呼和浩特 014010)

摘 要：多媒体信息由于维度高、数据量大、可解释性差等特征制约了其检索性能，提出了基于自然语言理解的智能化多媒体信息检索系统模型。该系统基于自然语言理解、数据挖掘、自反馈等技术的运用，在一定程度上扩大了检索范围，提高了检索准确率。

关键词：信息检索；自然语言理解；智能化；多媒体

中图分类号：TP393 **文献标识码**：A **文章编号**：1674-7720(2011)04-0006-05

Intelligent multimedia information retrieval system based on the natural language understanding

Shi Dongsheng

(Technology School of Information Engineering, Inner Mongolia University, Huhehaote 014010, China)

Abstract: The multimedia information retrieval performance was restricted because of higher dimensions, large amounts of data and bad explicable Features etc. In order to solve this problem, the intelligent multimedia information retrieval system model based on the natural language understanding was proposed. This system based on the natural language understanding, data mining, since the feedback technology in a certain extent, expanded search range, improving the retrieval accuracy。

Key words: information retrieval; natural language understanding; intelligent; multimedia

信息检索 IR(Information Retrieval)是指把用户所需信息按一定的方式组织起来的过程和技术[1]。传统的方式是用户通过输入关键字，从大量的文本库中检索出满足需求的文本，来判别文本是否相关并对相关文本进行排序的数学模型。然而随着网络的发展，信息资源不再以单一的纯文本传递为主，越来越多的信息资源以其他多媒体形式存储，如图像、视频、音频等，针对多媒体信息的检索近年来逐渐成为多媒体信息检索领域的研究热点[2]。参考文献[3]提出了基于本体信息检索系统的框架，该系统能够提取和利用网络上的语义信息，根据用户的检索条件进行推理，进而得出较为准确的结果；参考文献[4]提出了基于方法聚类的 Web 服务检索技术，该技术充分利用 Web 服务的描述信息生成基于方法层的 Web 服务建模方法，通过服务类聚算法产生基于方法层的服务检索模型及其相关算法；参考文献[5]提出了基于 Web 的智能信息采集处理系统，采用高效的 URL 去重和基于模版的下载机制，提高了采集 Web 资源的性能，并应用自然语言处理技术，对采集信息做智能分类和摘要，在发布上突出个性化的信息服务；参考文献[6][7]阐述了多媒体信息检索技术的发展现状。然而其研究仍存在以下不足：(1)搜索方式单一，信息相关性差；(2)不能准确地把握用户需求，容易产生搜索歧义；(3)搜索技术不具备智能化，搜索效率不高。为了解决上述问题，提出了基于自然语言理解的智能化多媒体信息检索系统 IMIRSTNLU(Intelligent Multimedia Information Retrieval System based on The Natural Language Understanding)。

1 IMIRSTNLU 模型概述

在该模型中，对多媒体信息的检索效果由词语分析和搜索服务共同决定，只有对多媒体信息词语分析准确，搜索服务才能够快速查找到与多媒体信息资源库中最贴近的资源，从而提供最贴近用户需求的多媒体信息。

该系统首先基于多媒体信息的资源分类，即通过对多媒体信息资源的自然理解，结合语言学和语义学学科

* 基金项目：内蒙古自然科学基金项目(20080404MS0910)

知识、专家知识及信息资源管理模式等，对多媒体信息资源在语义和知识层面上进行挖掘，训练成文本、视频、图像和音频四种常见格式的知识库[8]。

检索服务开始时，首先对用户输入的词语进行词语分析，挖掘出与用户输入词语相关度高的辅助语义，并提供给用户以确定最终检索语句。开始检索时，针对词语分析确定的语义条件，对知识库中的知识元采取相似度匹配方法，对多媒体信息的所有知识库启动二级搜索模式，即精确搜索和模糊搜索相结合。精确搜索某一模式知识库时，对另一模式知识库进行模糊搜索，若查找无结果，模糊搜索快速启动成为精确搜索，同时产生模糊搜索对未搜索知识库进行搜索。该方法针对用户输入词语进行词语分析，有效地提高了检索的准确率；对知识库的二级模式搜索，有效地提高了检索的效率。

检索结束后，对检索结果进行综合处理，去除无效链接、空链接及冗余数据等，依据与用户检索词语关联度的高低排列知识库中的资源记录，用户也可设定排列模式，如时间等。同时对检索情况的处理结果，如某一知识元按照用户检索习惯，应分类于哪一类知识库，更新多媒体信息资源的知识库。与此同时，保存用户的检索记录于用户资源列表，以便于下次检索生成更为确切的辅助语义。

2 IMIRSTNLU 定义

2.1 基础定义

定义 1 相似度匹配

数据以矩阵的形式存储于数据库表中，数据之间存在矩阵的相关性以及存储距离，因此根据不同形式的数据，其存储距离的大小不同，可以判定其相似度的大小。设数据信息 E 与 X 和 Y 的相似度为 P，则：

$$P_{X\to E}=\sum \min \| X-E \|^2 \quad (1)$$

$$P_{X\to F}=\sum \min \| Y-E \|^2 \quad (2)$$

$$P_{X\to E},\ P_{X\to F}\in P\cap P_A;\ P\in[0,1]$$

其中 P_E 的相似度为式(1)和式(2)的最小值，且 $P_E\in P[0,t_A]$，t 为知识库阈值。

定义 2 贴近度

若 P_E 的相似度值超过阈值 t_A，选择与之最贴近的阈值知识库进行相似度匹配。假设 $P_E>t_A$，且 $P_E<t_B<t_C<t_D$，则对知识库 B 进行搜索。

定义 3 词语分析

对词语经过解释处理，形成便于用户理解、有利于搜索的查询条件。设词语分析为M，则它包括 $M_{同义词分析}$、$M_{近义词分析}$、$M_{语义分析}$和 $M_{歧义分析}$ 4 个步骤。设数据信息 E，对其进行词语分析，首先会派生数据信息 E 关键词语相类同的多种信息，其中筛选与数据信息 E 的关键词描述意思相同的数据信息 $E_{同义}$，然后对其进行近义词分析，扩大数据信息 E 的查询范围，生成数据信息 $E_{近义}$，然后对数据信息 $E_{同义}$和 $E_{近义}$进行语义分析，筛选与搜索词语相贴近的数据信息 $E_{语义}$，最后经过歧义分析，形成搜索查询条件。

定义 4 辅助语义

在词语分析的基础上，根据用户使用习惯、个人兴趣爱好、搜索历史等条件对用户搜索查询条件给予一定的参考，帮助其提交合适、完善和更加准确的搜索查询条件。

3 模型介绍

该系统由以下几部分组成：(1)人机交互层。当用户输入检索词语后，系统提供相应的辅助语义提交给用户参考，用户确定满足实际需求的最终检索条件。信息检索结束后，搜索内容输出，显示给用户。(2)词语分析层。当用户输入搜索词语时，系统首先进行词语分析，对输入词语进行数据挖掘，分析与之相关联的数据信息，进行同义词分析、近义词分析、语义分析、歧义分析等，然后将挖掘的与之相关联的辅助语义推荐给用户，以供用户参考。(3)信息检索层。用户确定检索词语后，根据数据相似度匹配原则，启动精确搜索和模糊搜索相结合的模式，对多媒体信息资源知识库中满足检索条件的知识库记录进行查找。当相似度值确定后，属于某一知识库，即对该知识库启动精确搜索，同时启动模糊搜索对其余知识库进行搜索。如果搜索为空，则返回该搜索没有答案。否则输出该知识库中的信息记录。(4)搜索处理层。对搜索结果进行综合处理，去除无效链接，重复链接等，对信息的关键程度进行排序，保存搜索记录于知识库，并对知识库进行更新[9]，同时把用户的搜索习惯添加进用户习惯资源列表，以供下次搜索参考。通过不断收集用户搜索习惯和搜索结果，更新用户习惯资源列表和知识库，实现了用户个性化搜索。通过对搜索词语的自然理解解释，对知识库的动态更新、对搜索的二级模式设置、对辅助语义的记录等，实现了智能化，为以后快速定位搜索，创造了条件。具体框架图如图 1 所示。

4 算法分析

IMIRSTNLU 模型采用自然语言理解技术，结合数据挖掘方法，对用户搜索的数据信息进行检索。

知识库训练分类算法：

```
步骤(1)  A,B,C,D,K,E,P,T,M,N,L,p,t→0
                                  //变量初始化;
步骤(2)  scanf  E                 //输入信息 E;
步骤(3)  for (each E∈K)           //判别条件
    { if (p<tA)      //如果信息 E 与知识库 A 的相似
                     度在知识库相似度阈值范围内;
        then  E∈A; A=A ;printf ("A")
                     //把信息 E 添加进知识库 A,并更新;
      else  if (tA<p<tB)   //否则在知识库 B 的阈值范围内;
        then E∈B;B=B;printf("B")
                     //添加信息 E到知识库 B,并更新;
      else if (tB<p<tC)
```


《电子技术应用》 www.ChinaAET.com

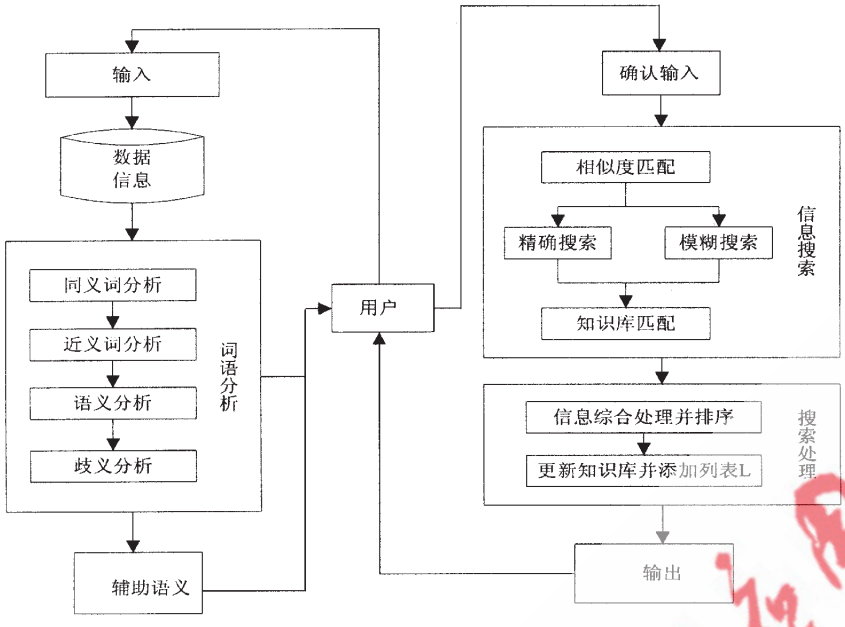


图 1 基于自然语言理解的智能化多媒体信息检索系统框架图

//否则,在知识库 C 阀值范围内;

then E ∈ C; C=C ; printf ("C")

//添加信息 E 到知识库 C,并更新;

else E ∈ D; D=D; printf ("D")

//添加信息 E 到知识库 D,并更新;

End if} End for

步骤(4) $K_A=A$; $K_B=B$; $K_C=C$; $K_D=D$;

//对知识库 K 进行更新存储;

多媒体信息检索算法:

步骤(1) A, B, C, D, K, E, P, T, M, N, L, p, t0

// 变量初始化;

步骤(2) $e \rightarrow M_{同义词分析} \rightarrow M_{近义词分析} \rightarrow M_{语义分析} \rightarrow M_{歧义分析} N$

//词语分析,生成辅助语义;

步骤(3) $e \rightarrow (M \cup N) \rightarrow E$ //用户确定搜索词语;

步骤(4) for (search E ∈ K)

{if ($p<t_A$) //数据信息 E 与知识库 A 的相似度小于知识库 A 规定的阀值 t_A;

$T_1=0$ //启动精确搜索;

Scan A //扫描知识库 A;

$T_2=1$ //启动模糊搜索;

Scan $K_{K \notin A}$

//扫描知识库 K 中除知识库 A 外的知识库;

if Scan==null; break "not answer"

//如果搜索为空,返回;

else printf search $E=A_E$

//输出知识库 A 中与搜索相关的数据;

$A=A_E$ //知识库 A 更新;

else if ($t_A<p<t_B$)

$T_2=0$ //模糊搜索更新为精确搜索

Scan B //扫描知识库 B;

$T_3=1$ //启动模糊搜索;

Scan $K_{K \notin (A \cup B)}$

//扫描知识库 K 中除知识库 A 和 B 外的知识库;

if Scan==null; break "not answer"

//如果搜索为空,返回;

else printf search $E=B_E$

//输出知识库 B 中与搜索相关的数据;

$B=B_E$ //知识库 B 更新;

else if ($t_B<p<t_C$)

$T_3=0$ //模糊搜索更新为精确搜索

Scan C //扫描知识库 C;

$T_4=1$ //启动模糊搜索;

Scan $K_{K \notin (A \cup B \cup C)}$

//扫描知识库 K 中除知识库 A、B 和 C 外的知识库;

if Scan= = null ; break "not answer"

//如果搜索为空,返回;

Else printf search $E=C_E$

//输出知识库 C 中与搜索相关的数据;

$C=C_E$ //知识库 C 更新;

else if ($t_C<p<t_D$)

$T_4=0$ //模糊搜索更新为精确搜索;

Scan D //扫描知识库 D;

if Scan==null; break "not answer"

//如果搜索为空,返回;

Else printf search $E=D_E$

//输出知识库 D 中与搜索相关的数据;

$D=D_E$ / /知识库 D 更新;

End if }; End for

步骤(5) N→L;

//辅助语义添加到用户习惯资源列表队列;

算法流程图如图 2 所示。

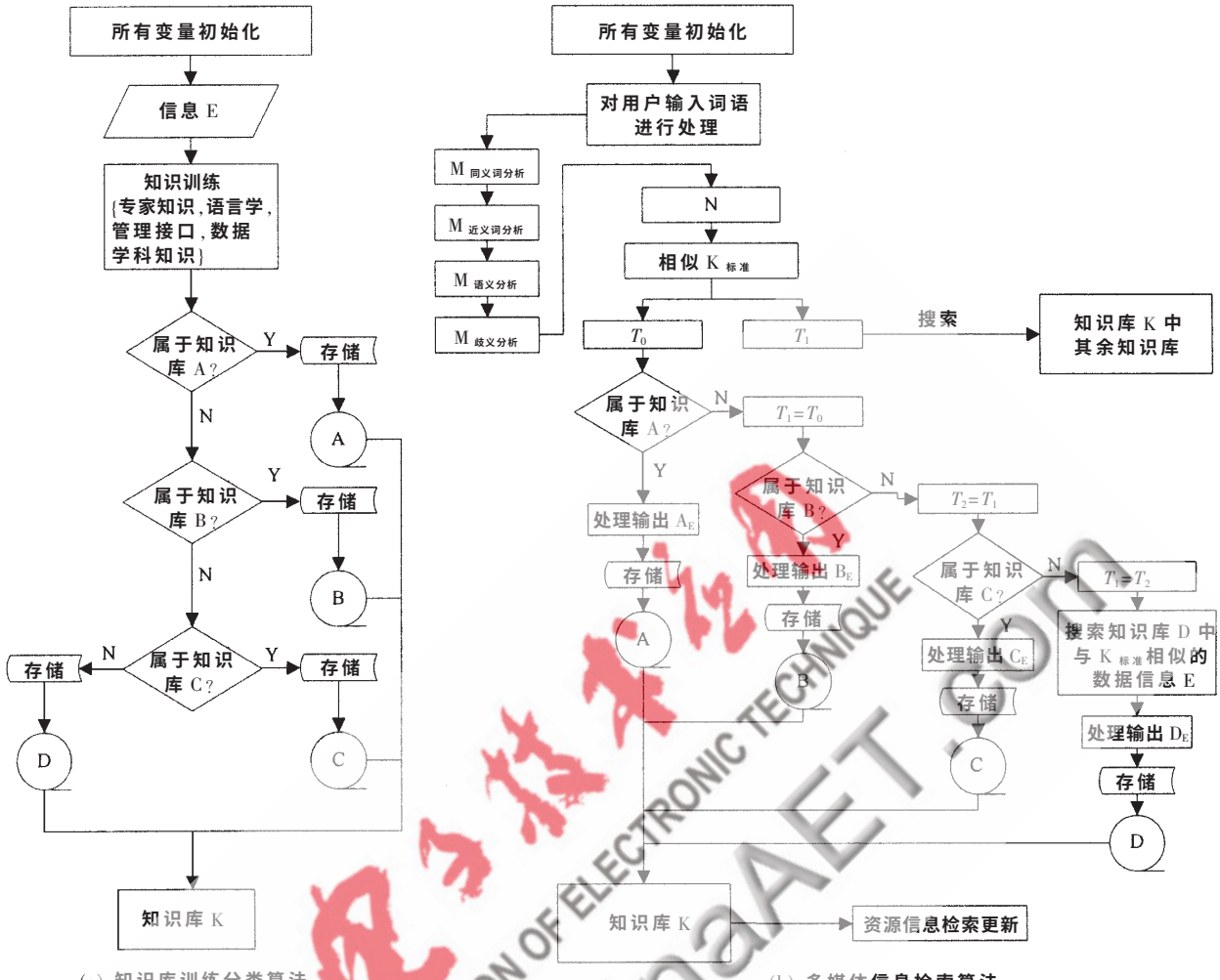
5 性能分析

由于目前针对多媒体信息检索研究还没有公认的数据集,所以本实验设计的数据库为文本、音频、视频和图像各 10 000 份所组成的实验数据库。实验平台为服务器一台 IBM3650, 基本配置为 2×4 core 2 GB CPU; 8 GB 内存; 500 GB 硬盘; 操作系统为 WIN2003 SERVER 标准版; 编程环境为 VC++2005。由于事先设定了各知识库的文件数量,所以知识库的组成已经得知,如表 1 所示。

表 1 IMIRSTNLU 系统知识库组成

	文本	视频	音频	图像	知识库
知识元素	10 000	10 000	10 000	10 000	40 000

对实验结果的评测,采取信息检索中常用的三个指标: 检全率 Recall、检准率 Precision 和 F1-measure 值,其定义如下:



(a) 知识库训练分类算法　　(b) 多媒体信息检索算法

图 2 IMIRSTNLU 模型算法流程图

$$\text{Recall}=\frac{I}{W}\times 100\%,\ \text{Precision}=\frac{I}{R}\times 100\%,\ \text{F1-measure}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

其中 I 为检索到的满足检索方法的检索数，R 为检索结果数，W 为可供选择的检索数。实验时分别输入针对 4 种知识库检索的检索条件，经由 IMIRSTNLU 系统对其进行搜索，经过式(3)、式(4)和式(5)对实验数据进行处理计算，结果如表 2 所示。

同时该实验对多媒体信息检索的效果与参考文献[10]的检索效果进行了对比，具体如图 3 所示。其中星号表示该实验的 F1-measure 值，圆圈表示参考文献[10]的 F1-measure，通过对比可知，该系统的检准率与参考文献[10]相比有明显的提高，能够基本实现智能化理解用户检索需求，同时由综合评价 F1-measure 值可以看到，该系统的检索服务是高效和准确的。

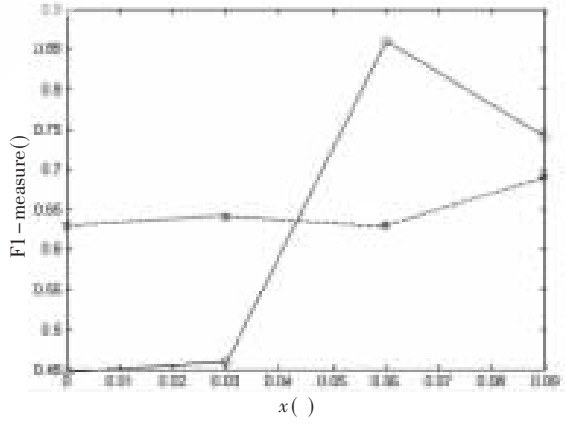


图 3 该模型检索效果与参考文献[10]的效果对比图

表 2 IMIRSTNLU 系统实验数据分析

	W/条	R/条	I/条	Recall/%	Precision/%	F1-measure()
A	10 000	9 402	6 741	67.41	71.70	0.69
B	10 000	9 391	6 198	61.98	66.00	0.63
C	10 000	9 643	6 304	63.04	65.37	0.64
D	10 000	9 598	6 125	61.25	63.82	0.63

本文经过对自然语言和数据挖掘技术的理解，

提出了一种智能化多媒体信息检索系统，通过对用户输入词语进行词语分析，生成辅助语义帮助用户参考搜索查询条件，启动二级模式搜索，对知识库实现全面和准确的搜索，同时对搜索结果进行综合处理，对知识库实现不断更新，对用户使用习惯进行存储记忆，有效地解决了检索语义模糊不清，查找范围不全和准确率不高的问题。

参考文献

[1] Liu Ying, Tang Yonglin, Zeng Yuan. A study on improving information retrieval effectiveness for scientific and technical novelty retrieval[C]. Proceedings of International Forum on Technological Innovation and Competitive Technical Intelligence '2008, 2008:338-347.

[2] JAIN P. Intelligent information retrieval[C].SETIT 2005 3rd International Conference:Sciences of Electronic, Technologies of Information and Telecommunications, 2005,3:27-31.

[3] KANNAN R. Topic map: an ontology framework for information Retrieval[C]. Proc.of National Conference on Advances in Knowledge Management 2010:195-198.

[4] Peng Dunlu, Zhou Aoying. Web service retrieval technology based on the method of clustering[J].Computer Applications, 2007,27(10):2365-2368.

[5] Zhang Fan, Li Linna, Yang Bingru. The intelligent information collection and processing system design and implementation based on the Web [J]. Computer Engineering, 2007, 33(18):265-267.

[6] GOYAL P, BEHERA L, MCGINNITY T M. Application of bayesian framework in natural language understanding [J]. IETE Tech Rev, 2008, 25(5):251-269.

[7] TANENHAUS M K, SARAH B S. Language processing in the natural world[J]. Phil's Trans R Soc Lund B Boil Sic. 2008, 363(1493):1105-1122.

[8] LEE C, LEE G, JANG M.Dependency structure language model for information retrieval[C].ETRI,2006,28(3):337-346.

[9] CAO G, NIE J, BAI J. Integrating word relationships into language models[C].Proc.28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.Brazil. 2005:298-305.

[10] Liu Wei, Chen Junjie. A framework for intelligent meta-search Engine Based on Agent[J]. Computer Engineering end Application, 2005, 3:137-211.

(收稿日期:2010-11-25)

作者简介：

师东生，男，1970年生，讲师，硕士研究生，主要研究方向：数据挖掘、图像处理、模式识别。

APPLICATION OF ELECTRONIC TECHNIQUE www.chinaAET.com