面向音乐领域的文本检索与挖掘系统*

付瑞吉，秦兵，刘挺
哈尔滨工业大学计算机学院信息检索研究中心 哈尔滨 150001
Email: {rjfu, bqin, tliu}@ir.hit.edu.cn

摘 要：本文介绍了一个面向音乐领域的文本检索与挖掘系统——八维音乐资讯，主要通过信息抽取技术，对音乐领域的大量半结构化和非结构化的文本进行深层次的挖掘，从中抽取出有价值的信息，转换为结构化数据，目的是为用户提供精准化、全方位的音乐检索结果。该系统从人、物、时、空、评价、数量、事件和关系八个角度出发，集成了命名实体识别、关系抽取、事件抽取、倾向性分析、问答等多项自然语言处理和信息抽取技术。系统设计采用 MVC 架构，包括 3 层结构：模型层、控制层和视图层。总之，该系统将已有的信息抽取技术应用于音乐垂直检索系统，具有一定的新颖性和实际意义。

关键词：八维音乐，信息抽取，垂直搜索

Text Retrieval and Mining System for Music Domain

Fu Ruiji, Qin Bing, Liu Ting
Research Center for Information Retrieval of Computer Science & Technology School, Harbin Institute of Technology, Harbin 150001)
Email: {rjfu, bqin, tliu}@ir.hit.edu.cn

Abstract: This paper presents a text retrieval and mining system for music domain, named 8wei Music Information. By means of information extraction (IE), it can mine a great deal of semi-structured and unstructured text deeply, extract valuable information from it and convert the information into structured data, which aims to return accurate and overall search results to users. From persons, objects, time, space, opinions, quantity, events and relations 8 standpoints, 8wei Music system integrates many kinds of natural language processing (NLP) technologies and IE technologies, such as named entity recognition, relation extraction, event extraction, sentiment classification, question-answering, etc. The MVC software framework is used in the system, which consists of three components: Model, View and Controller. In brief, existing IE technologies are used for music vertical search system, which has a certain novelty and practical significance.

Keywords: 8wei Music; Information Extraction; Vertical Search

1 引言

随着互联网技术的发展，尤其是进入 web2.0 时代以来，博客、RSS、WIKI、SNS 等社会软件的涌现，使每一个用户都可以成为信息的发布者，网络上的信息迅猛增长。因此我们迫切需要一些自动化的工具帮助人们在海量信息源中迅速找到真正需要的信息。在这个背景下产生了信息检索（Information Retrieval，IR）技术和信息抽取（Information Extraction，IE）技术。

信息检索是指将信息按照一定的方式组织和存储起来，并根据用户的需要找出相关信息的过程[1]。目前成功的信息检索系统有 Google，百度，Yahoo!，Bing 等著名的搜索引擎，但这些都是通用搜索，返回的信息过于繁杂，噪音很大，这极大地增加了用户甄别信息价值的时间，并不能满足特殊用户群、特殊领域的精准化信息服务需求。于是人们开始关注垂直搜索引擎，针对某一个领域进行精准、细致、全面的搜索。这就需要信息抽取技术[2]，从非结构化的文本中抽取出特定的信息，对海量的信息进行精准全面的挖掘。

八维音乐资讯[1]就是一个以信息抽取为基础的音乐领域的垂直搜索引擎系统。八维指的是描

* 基金资助：国家自然科学基金项目（60803093，60975055）；国家 863 项目（2008AA01Z144）
作者简介：付瑞吉(1984-)，男，陕西省府谷县人，博士研究生，信息抽取；秦兵（1968- ），女，教授，博士生导师，信息抽取和多文档文摘；刘挺（1972- ），男，教授，博士生导师，信息检索和自然语言处理。

述事物的八个维度，包括：人、物、时、空、评价、数量、事件和关系。系统集成了命名实体识别、关系抽取、事件抽取、倾向性分析、问答等多项自然语言处理和信息抽取技术，旨在为用户提供更加精准和全面的搜索结果。

2 系统描述

八维音乐资讯提供了音乐资讯检索，关系抽取，事件抽取，倾向性分析和问答等功能。

2.1 音乐资讯检索

八维音乐资讯提供音乐领域的资讯检索服务，用户可以输入感兴趣的 query，系统返回相关的音乐资讯，包括标题、内容片段、URL 等信息。支持分时段检索，用户可以选择资讯的发布时段，检索该时段内和 query 相匹配的资讯信息。

2.2 关系抽取

八维音乐资讯自动地从音乐资讯网页中抽取音乐实体之间的关系，目前包括“艺术家-歌曲”、“艺术家-专辑”、“艺术家-唱片公司”、“歌曲-专辑”、“专辑-发行时间”5 种关系，供用户浏览和查询。

2.3 事件抽取

八维音乐资讯自动地从音乐资讯网页中抽取音乐事件，目前包括艺术家举行演唱会和发行专辑 2 类事件，抽取事件的时间、地点、人物等信息，可供用户查询艺术家的相关动态，也可以向用户提供演唱会预告和新专辑预告等服务。

2.4 倾向性分析

八维音乐资讯自动地从音乐评论网页中抽取网友对于音乐的评论，包括评价对象的属性和评价词，可供用户查询和浏览音乐实体的相关评论，也可根据评价词查找相符的音乐实体。

2.5 问答

八维音乐资讯提供了简单的一次交互问答服务，当用户输入一个问题时，系统经过问题分析并查询相关的数据库，返回精准的答案。可处理的问题类型包括关系型、事件型、评论型等。

3 系统框架

3.1 设计思想

八维音乐资讯框架的设计思想主要包括两点：

1) 借鉴 MVC[3]分层设计，从下到上分为模型层、控制层和显示层，层与层之间逻辑独立，通过数据传输进行连接；
2) 将底层模块封装成网络服务（Web Service），各服务模块尽量相互独立，便于增删改和并行开发。

八维音乐资讯的系统使用基于 python 语言的 Django[2]作为开发框架。系统框架如图 3-1 所示，其中的箭头表示数据的流向。

---

[1] http://yue.8wss.com
[2] http://www.djangoproject.com/

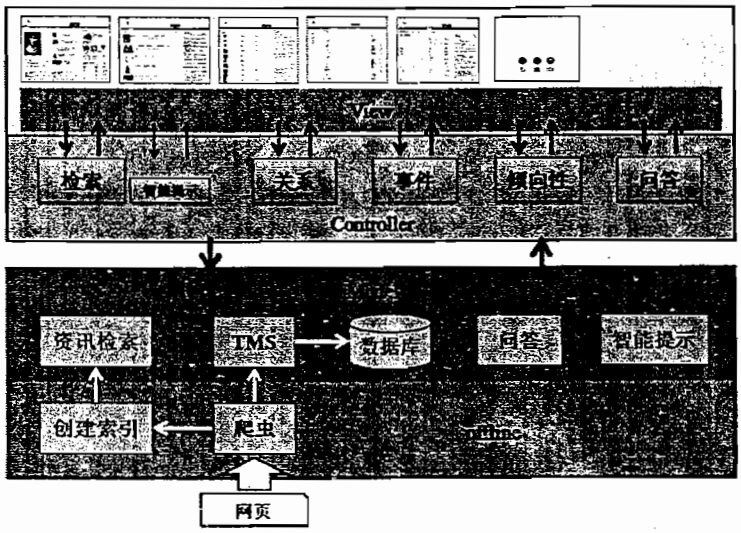

图 3-1 八维音乐资讯系统框架图

模型层（Model）：

模型层包括数据库以及各底层模块的网络服务，包括资讯检索服务，文本挖掘系统（TMS）服务，问答服务，智能提示服务以及离线部分（offline）。

离线部分又包括爬虫模块和创建索引模块。爬虫程序定时爬取网页，然后一方面调用文本挖掘系统（TMS）服务处理这些网页，并将信息抽取的结果保存到数据库中；另一方面对网页创建索引，供资讯搜索引擎使用。

控制层（Controller）：

处理客户端发起的各种请求，调用模型层的服务或数据库，返回相应的结果。每一个功能对应一个模块，模块之间相互独立，负责处理相应的请求。

显示层（View）：

系统与用户交互的接口，负责将数据以网页的形式显示出来，同时提供数据输入的接口。

下面我们将重点对模型层的主要服务模块做详细介绍。

3.2 服务模块

3.2.1 资讯检索服务

我们使用了基于 Lucene[4]的搜索引擎技术，对于爬虫定时爬取到的音乐资讯网页，进行正文提取和分词等处理，然后建立全文倒排索引，并且按照时间排序，便于推送最新资讯和分时段检索。目前，我们爬取网页和更新索引的频率是每小时一次，以保证用户能够搜索到最新的资讯。

我们将资讯搜索部署为若干个网络服务，由系统的控制层负责调用，控制层从视图层接收信息，并转换为相应的网络访问地址，通过 http 请求调用服务。这样控制层就能将接收到的用户的查询封装成 http 数据报文，发送给网络地址对应的资讯搜索服务程序，服务程序解析报文，提取查询，在索引中进行检索，得到结果，同样以报文形式返回给控制层，控制层再解析出检索结果，发送到视图层，显示给用户。

3.2.2 文本挖掘系统

文本挖掘系统（Text Mining System，TMS）是八维音乐资讯的核心部分，主要完成信息抽取的功能，包括底层自然语言处理部分以及关系抽取模块、事件抽取模块、共指消解模块和倾向性分析模块。TMS 负责对输入的文本进行深层次的挖掘，抽取我们感兴趣的信息，实现将非结构化或半结构化的文本转换为结构化的信息，处理的结果存储在数据库中。

图3-2显示了TMS的数据模型层的处理流程图。当输入一个文本时，TMS首先会调用底层的NLP处理工具对文本进行处理，包括断句、分词、词性标注和命名实体识别等，然后进行文本挖掘的上层处理，包括关系抽取、事件抽取、共指消解和倾向性分析，处理的信息都保存在一棵DOM树上，便于使用。

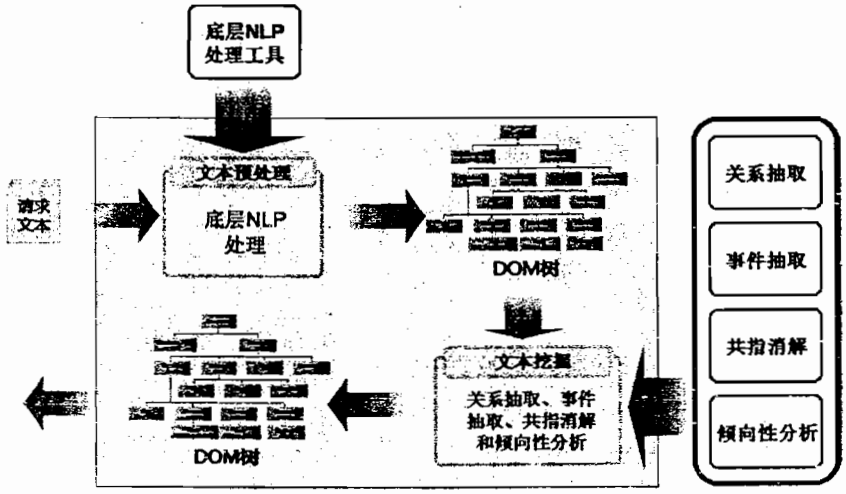

图3-2 TMS数据模型层处理流程

1）命名实体识别

命名实体(Named Entity，NE)识别的任务是识别出文本中特定的实体，它是信息抽取、机器翻译、信息检索和自动问答等多种自然语言处理技术的重要基础。我们目前的研究不但包括通用域命名实体的识别，还扩展了音乐领域的实体研究。扩展的实体包括音乐名、艺术家名、专辑名和艺术家别名4类。

目前通用域命名实体识别主要有规则和统计两类方法。统计方法的健壮性和灵活性更好，可以更方便地在不同领域之间移植，我们在人民日报语料上，采用最大熵马尔科夫模型（Maximum Entropy Markov Model，MEMM）方法，获得了90%以上的识别准确率。

同时，我们根据音乐命名实体的特点，在语料库规模有限、语料质量较差的条件下，提出一种规则和统计模型相结合的分类方法[5]来进行音乐NE的识别。首先通过音乐专业词典以及规则匹配出音乐NE的候选，然后通过最大熵（MaxEnt）模型进行分类，完成音乐NE的识别。

为了更加直观地展示这种分类方法的特点，我们将它和词典匹配方法、基于隐马尔可夫模型（Hidden Markov Model，HMM）的序列标注方法以及基于最大熵马尔可夫模型的序列标注方法进行了比较实验。实验证明我们的分类方法音乐NE识别性能最好。

总的来说，我们提出的分类方法能够使用更加丰富的特征，更加有效地利用音乐词典的信息，克服了语料库规模较小、质量较差的困难，达到了较好的音乐NE识别效果，精确率、召回率和F值分别达到了89.89%，81.01%，87.93%。此外，相对于MEMM序列标注，该方法还有模型小，运行效率高的优点。

2）实体关系抽取

实体关系抽取的任务是从文本中抽取出两个或者多个实体之间预先定义好的实体关系。我们将实体关系抽取定义为一个分类问题，主要研究内容是中文音乐领域的实体关系抽取。针对这一问题，本文首先构建了中文音乐实体关系语料库，然后分别采用了基于序列模式挖掘的无指导的方法和基于特征提取的有指导的方法来解决这一问题[6]。

在语料库的建设过程中，我们参考了ACE语料的构建过程，首先定义了包含5种实体关系的中文音乐领域关系类型体系；接着制定了详细的标注规范并完成了10,000句语料的标注工作。

与此同时，针对音乐领域和中文的语言学特点，定义了音乐领域的序列模式。由于 BootStrapping 方法的引入，实体关系种子可自动扩展并可从互联网上挖掘大量的高准确率的序列模式。在评测集上，该方法取得了平均准确率为 94.40%的结果，但是召回率很低。在互联网海量语料的前提下，具有一定的实用性。同时我们在基于已标注完成的语料库基础上，研究了音乐领域实体关系抽取的特点，并根据其特点进行了特征选择的研究，分别使用最大熵和支撑向量机（SVM）对特征抽取的结果进行了实验，在相同的测试集上，SVM 分类器取得了更好效果。目前，我们采用混合核的技术，使音乐领域的实体关系准确率和召回率都有提高，F 值达到 82%以上。

3）事件抽取及动态关系维护

事件抽取是信息抽取领域一个重要的研究方向。事件抽取主要把人们感兴趣的，用自然语言表达的事件以结构化的形式呈现出来，如什么人，什么地方，什么时间，做了什么事等，在自动文摘，自动问答以及信息检索等领域有着广泛的应用。

我们首先在 ACE 语料上，分析和研究事件抽取方法的研究，提出了基于触发词扩展的事件类型识别和事件元素识别的方法，取得了一定的结果。同时，我们从音乐领域切入，选择了具有代表性的演唱会及专辑事件进行深入研究。首先借鉴 ACE 评测中事件抽取任务的相关概念以及构建语料库的一些经验，详细定义了我们所关注的两类事件，并且构建了语料库，并给出了语料标注的来源、过程、标注规范以及存储格式等。

对事件抽取的两项关键技术——事件类型识别以及事件元素识别采用不同的处理策略。事件类型的识别采用了基于关键词与触发词相结合的过滤方法[7]。

在事件元素识别中，如何从众多的实体中找出事件元素，成为我们研究的重点。我们提出了两种方法：基于模式匹配的事件元素识别以及基于最大熵的事件元素识别。在总结前人事件表示模型的基础上，我们结合汉语的特点以及所采用句法分析模块的特点提出了一种基于简化依存句法树模式匹配的方法；基于最大熵的方法将事件元素识别问题看作分类问题，将所有出现的实体作为候选事件元素，选取上下文、邻近实体、句法结构等特征从不同的角度描述候选元素，并采用最大熵分类器对其进行二元分类。为了发挥各自方法的优点，将基于模式匹配的方法与基于最大熵分类的方法采用级联的方式形成最终事件元素识别的解决方案。在目前音乐领域典型事件类型下，事件识别的平均 F 值达到 83.84%，事件元素识别的平均 F 值达到 76.41%。

4）共指消解

共指是一种特殊的实体关系，共指消解目标是抽取篇章内的实体的等价关系。共指消解本身是一个非常复杂的问题，需要考虑的问题和因素很多。共指消解的本质是等价类划分。根据对划分过程的影响和处理策略，我们针对共指消解研究中多个层面的问题进行了深入的研究，主要是在一些共性的问题上进行了探索。

在共指消解领域，很多传统特征都被证明是非常有效的。近来，越来越多的研究人员开始针对共指消解挖掘更加丰富的特征，例如各种背景语义特征。我们通过挖掘基于 WordNet、HowNet、维基百科和浅层语义知识等多种背景语义知识，进行有效特征的选择，构建共指模板特征，很好的表达特定的语义关系。在进行特征选择时采用最大熵，确定最优的特征组合后使用 SVM 进行重新训练和测试。实验结果表明，基于背景语义知识的系统提高了性能指标。

我们针对音乐领域对共指类型进行细分类，分为第三人称代词的指代、指示性代词修饰的名词短语的指代以及例如“这”、“那”等这类代词的指代。我们把每个类别的解决看作共指消解的一个子问题，首先对 Mention 进行分类，对不同类别的指代采用不同的策略。在经过不同消解器进行消解之后得到的都是二元指代对，还需要对这些指代对进行合并，直到没有冲突或者超出距离范围（对照应语和先行词的距离进行了限制）。我们采用三种目前国际上普遍采用的共指消解评测方法对系统进行了详细地评测，评测结果的平均 F 值达到了 90%以上[8]。

5）倾向性分析

倾向性分析主要是针对主观性文本单元自动获取有价值的意见信息，是一个新颖且非常有应用价值的研究课题。倾向性分析技术可被广泛应用于多种自然语言处理问题中[9]。八维音乐资讯中倾向性分析的目的是为音乐评论中的音乐实体打上一些情感标签。这些情感标签多使用情感词来体现，如为音乐实体"周杰伦"打上"有个性"，"率真"等情感标签。对于用户而言，可以通过简洁的标签快速了解该音乐实体的一些特性，省去用户大量搜索和阅读相关评论的时间。

我们通过对音乐评论文本的分析，挖掘出评论文本中的<音乐实体，标签>二元对。为此，我们构建了极性词、维度词和单一标签库等词表，用于评价元素的识别和抽取。首先，我们对原始的评价文本进行预处理，包括分词、词性标注、音乐实体识别的结果修正，如错误分词合并、词性修正、短语合并、标签符号修正等等。然后以词表为基础，按照相应的算法流程对音乐评论文本进行音乐实体标签抽取，得到关于音乐评论文本的一个初步的<音乐实体，标签>集合。音乐实体标签挖掘部分可以挖掘出音乐评论文本中大部分的标签，但还有一些特殊的标签难以被发现；此外，结果中可能还存在一些错误标签。基于此，我们使用一些规则来对音乐实体标签进行一定的修正，以提高标签挖掘模块的性能。

经过评测，本模块的准确率达到了76%以上，召回率达到80%以上。

3.2.3 问答服务

问答(Question Answering，QA)研究的目标是为使用者提供更加自然的信息访问交互界面[10]。面对输入的一个问题，问答系统首先要对问题进行分析，明确问题对预期答案的语义约束条件。通常有经验规则和统计机器学习两大类方法。

目前对八维音乐资讯的问题分析暂时采用简单的基于规则的处理策略。对于一个用户的查询，首先使用TMS处理，然后在此基础上判断其是否为问题，利用的主要特征为是否含有疑问词。如果不是问题，返回非问题提示；如果是问题，则对该问题进行进一步分析。

我们根据当前八维音乐资讯数据库的特点，将问题分为关系、事件、倾向性和其他四类。前三类问题分别针对该类问题的特点分析问题询问的焦点项，并把分析结果返回，供控制层进行数据库查询以获取精准的答案。

```
<YTQA>
    <TYPE>relation</TYPE>
    <SUBTYPE>NaNc</SUBTYPE>
    <Na>周杰伦</Na>
    <Nc>歌曲</Nc>
    <FOCUS>Nc</FOCUS>
</YTQA>
```

图3-3 问答服务处理结果示例

为准确获取关系、事件、倾向性三类问题的焦点项，我们采用基于规则的方法。对这三类问题结合各自数据库表结构的特点进行了详细分析，在此基础上对三大类问题还划分了若干子类，并确定问题分类和分析的特征，为每一子类问题都确定了该类问题的元素向量，如对于关系类型中的"艺术家-歌曲（NaNc）"问题，其元素向量便为(Na, Nc)。每一类元素都有特殊值和一般值，如"周杰伦"、"范玮琪"是Na元素的特殊值，"歌手"、"歌星"是Na元素的一般值，"歌曲"是Nc元素的一般值。为此，我们构建了每一类元素的一般值词表。利用TMS对用户query的分析结果和词表，可以确定查询中具备哪些元素的哪些值，而值为一般值的元素便为问题的焦点所在，从而最终确定了问题的焦点项。例如：用户查询为"周杰伦唱过什么歌曲"，TMS的分析结果为"周杰伦/Na 唱/v 过/u 什么/r 歌曲/n"，再加上词表的信息可以确定该查询含有Na、Nc

两个元素，且 Na 具有特殊值“周杰伦”，Nc 具有一般值“歌曲”，所以该查询为关系类型中的“艺术家-歌曲（NaNc）”问题，其问题焦点是 Nc。问题分类结果的形式如图 3-3 所示。

特别地，如果查询中含有的所有元素的值都为特殊值，则该问题的焦点项为“是否”。例如查询为“《青花瓷》是周杰伦唱的吗”。这种问题需要根据各元素的值在对应的数据库中验证记录是否存在，返回“是”或“否”。

目前的方法在准确率上取得了不错的效果，但召回率不高，而且可移植性较弱。在今后的工作中，我们希望在规则的基础上，可以发现一些分类能力强的特征，再引入机器学习的方法进行问题分类。

4 结论

本文介绍了一个面向音乐领域的文本检索与挖掘系统。该系统提供了资讯检索、关系抽取、事件抽取、倾向性分析和问答等服务功能，为用户提供精准全面的信息。系统架构划分为模型层、控制层和视图层。其中模型层包括资讯检索模块、文本挖掘系统（TMS）、问答服务和数据库等部分，采用低耦合的网络服务的设计思想，各部分相互独立，便于改进更新。

TMS 是系统的核心部分，它集成了命名实体识别、实体关系抽取、事件抽取、共指消解、倾向性分析等多个自然语言处理和信息抽取模块，负责对输入的文本进行深层次的挖掘，将非结构化或半结构化的文本转换为结构化的信息，为精准化检索和问答等服务做好了数据准备。问答服务也是基于 TMS 对用户查询的处理结果，对查询进行问题判别与分类，然后到数据库中查找对应的答案。

5 下一步工作

在下一步的工作中，我们将不断改进底层模块的性能，提高信息抽取的准确率和召回率，将更加精准全面的信息展现给用户。同时，我们还将优化系统的架构，提高服务的速度，增强系统的稳定性。另外，用户的需求作为技术发展的指南针和推动力，也很重要，因此我们将继续调研用户的需求和使用习惯，不断改进系统的功能以及呈现方式。

另一方面，我们的研究不仅仅局限在音乐领域，如何能够方便迅速地迁移到其他领域，也是我们的研究兴趣点之一。

参 考 文 献

[1] Ricardo Baeza-Yates, Berthier Ribeiro-Neto. 现代信息检索[M]. 北京:机械工业出版, 2004:1-17.

[2] 李保利, 陈玉忠, 俞士汶. 信息抽取研究综述[J]. 计算机工程与应用, 2003, 39(10):1-5.

[3] 闫宏印, 张卫争, 刘超慧. 开源框架下 Web 应用分层的设计与实现[J]. 计算机工程与设计, 2008, 29(23): 6023-6028.

[4] Otis Gospodnetic, Eric Hatcher 著, 谭鸿, 黎俊洪等译. Lucene IN ACTION 中文版[M]. 北京:电子工业出版社, 2007:1-148.

[5] 付瑞吉, 车万翔, 刘挺. 一种基于分类方法的音乐命名实体识别技术[J]. 黑龙江大学自然科学学报, 2009,26(增刊):62-69.

[6] 周蓝珺. 音乐领域中文实体关系抽取研究[D]. 哈尔滨:哈尔滨工业大学, 2009.

[7] 宋凡. 音乐领域典型事件抽取技术的研究[D]. 哈尔滨:哈尔滨工业大学, 2009.

[8] 郎君, 忻舟, 秦兵, 刘挺, 李生. 集成多种背景语义知识的共指消解[J]. 中文信息学报, 2009, 23(3):3-9.

[9] 刘鸿宇, 赵妍妍, 秦兵, 刘挺. 评价对象抽取及其倾向性分析[J]. 中文信息学报, 2010, 24(1):84-93.

[10] 张志昌, 张宇, 刘挺, 李生. 开放域问答技术研究进展[J]. 电子学报, 2009, 37(5):1058-1069.