

(10) 国际公布号
WO 2015/043075 A1

- (81) 指定国** (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR,

[见续页]



CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

- (84) **指定国** (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,

AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

- 包括国际检索报告(条约第 21 条(3))。

本发明是一种面向微博的情感实体搜索系统。包括以下 5 个模块: 1) 用户接口, 用于系统与用户的交互, 用户可以通过该模块提交查询请求并获得反馈结果; 2) 查询扩展模块, 用于对微博语料数据进行词语关系挖掘, 并结合 Word Net 本体库建立加权词语关系图; 3) 查询处理模块, 用于将用户查询请求转换为索引库所能接受的查询关键词及查询语句, 并基于模块 2) 构建的词语关系图进行查询扩展; 4) 情感信息挖掘模块, 用于对微博语料库进行情感挖掘, 并生成情感实体及情感极性的判定规则; 5) 情感信息判定及索引建立模块, 用于对微博数据进行情感实体和情感极性的判定, 建立情感信息索引, 并进行存储; 6) 倒排索引建立模块, 用于对微博文本信息建立倒排索引, 并进行存储。本发明解决微博情感实体抽取、情感极性分析和情感实体搜索等困难问题, 为社交网络舆情分析和监控提供一种智能搜索产品。

面向微博的情感实体搜索系统

技术领域

本发明涉及文本情感挖掘及信息检索领域，具体涉及一种面向微博的情感实体搜索系统，属于面向微博的情感实体搜索系统的创新技术。

背景技术

近年来，随着互联网和社交网络的发展，包括微博在内的社交网络数据正以指数形式快速增加。微博的不断增长使得人们可检索的信息越来越丰富，但是海量的微博数据也使得人们难以迅速、准确地找到所需要的信息。同时，由于微博行文上的自由性，情感信息的抽取相对于传统文本更为困难，在对舆情监控和产品调研行业具有重要意义的微博情感信息检索领域，目前还未见成熟的技术和系统。

面向微博的情感实体搜索方法和系统主要涉及三类相关的关键背景技术。其一为查询扩展技术；其二为情感实体抽取技术；其三为情感极性判别技术。以下分别对上述三类背景技术分别加以阐述和分析。

1 查询扩展技术

直接通过关键词进行查询的传统检索系统或搜索引擎可以获得一些相关的检索结果，但是这种利用简单匹配的方式查找的结果较为机械，不能真正理解用户的查询意图，返回的结果也就无法让人满意。因此寻找一种

方法可以很好的理解用户的查询意图，提高检索的查准率和查全率成为解决上述问题的热点。查询扩展技术正是这样的一种方法。通过查询扩展可以更准确的理解用户查询需求，帮助用户更快更准确地获得需要的信息。经典的查询扩展方法主要包括基于全局分析、基于局部分析、基于用户查询日志和基于关联规则四种。近年，有学者提出基于本体(或领域本体)和语义网的查询扩展方法。

基于全局分析的查询扩展方法是通过挖掘全部数据集或整个数据库的文档中的词语相关度进行扩展的。其优点在于可以对整个数据集进行充分的分析，能够了解文档的方方面面；其缺点是，由于通常的数据集都过大，因此对分析的时间和设备的要求都很高，更不可能在线完成。现有的检索系统都是在离线完成全局词语的分析，对于需求实时的搜索引擎更是难以采用这种方法。

基于局部分析的方法包括相关反馈与伪相关反馈两种。相关反馈即是先通过用户初始查询，得到检索结果，然后再由用户人工判断结果文档的相关与不相关，分放于两个不同的文档集。这样就获得了标记的相关文档，作查询扩展前只需要对这些文档进行词语分析即可。这样做的优点是只处理相关部分的文档，使得文档数量减少了，而且相关度也会有所提升；其缺点是需要大量的人工反馈，这需要大量的人力，而且仍然需要大量的实验进行调试处理。这样现有的检索系统或搜索引擎少有采用这种方法的。

伪相关反馈方法是利用用户初次查询获得的前 n 篇结果进行分析，其理论假设是认为结果中与查询词相关的文档会出现在检索的最前面，也就是认为这些文档就是相关度最高的文档，通过分析这些文档获得扩展词并

进行查询扩展。专利申请号为 CN20091032193.5，发明名称为“查询扩展方法及查询扩展系统”就是利用伪相关反馈的专利例子。其主要思想是通过将用户初次查询所得结果靠前的部分文档通过聚类分析并生成簇，对簇进行排序后，再从排名在前一定数目的簇中提取扩展词，把所得的扩展词添加到原查询中，形成扩展词结合然后进行二次检索。这种方法的缺点是在于不能保证初次查询的靠前的文档就是相关的，如果是不相关的话，得出的扩展词可能会使得二次检索的结果更不相关，检索性能就会降低。

基于用户查询日志的方法是现在搜索引擎通用的一种扩展方法，该方法是通过用户对用户的查询日志进行词语分析，将共现的词语作为扩展词。专利申请号为 CN200710097501.6，发明名称为“查询扩展方法和装置以及相关检索词库”和专利申请号为 CN200810115470.7，发明名称为“一种扩展查询的方法、装置及搜索引擎系统”就是对用户输入的查询词进行分析得到相关的词语，然后将这些词语作为扩展词。这种扩展方法首先也需要获得大量的查询日志，这需要一个积累的过程。

基于关联规则的方法是一种数据挖掘的经典方法，常常是用于挖掘事务之间的相关性，在查询扩展中可以用于各种形式的资源进行挖掘，例如挖掘数据文档集、查询日志等资源的词语间的相关性。专利申请号为 CN201010605956.6，发明名称为“扩展用户搜索结果的方法及服务器”就是用关联规则技术进行查询扩展的例子。该专利采用一个关联规则数据库储存建立好的规则，这要的规则可以使手工建立也可以使利用支持度-置信度框架的关联规则对特定文档进行挖掘，将生产的规则保存到关联规则数据库中。当用户输入查询词时，首先在规则数据库中获取与该词相关的词，

然后将原查询词、获得的相关词以及两者的组合词形成新的查询词，并对数据库进行二次检索。这种方法的缺点是未能通过词语的意思层面去理解一个词语，只是浮在词语的频率层面上，这样的扩展也就不能很好的了解用户的查询意图。

基于本体或语义网的查询扩展方法通过使用或构建词语网络来对词语进行扩展的一种技术。这种语义网络可以是已经建立好的网络，如 WordNet 和 HowNet；也可以是自行构建的，如领域知识或领域本体。语义网或本体库组织了词语的多层关系，例如同位词、上下文位词、概念词、整体-部分词等等关系，使之组成一个关于词语的网络。专利申请号为 CN200810116729.X，发明名称为“一种基于领域知识的语义查询扩展方法”是先利用领域知识和用户语句特征的分析来构建一个领域知识库，然后利用领域知识库内容，对原查询词进行语义分析，获得一个语义项列表，再通过语义计算得到可扩展项；最后将扩展项返回查询集合中对数据库进行二次检索。专利申请号为 CN20101084725.2，发明名称为“一种图像检索中基于文本的查询扩展与排序方法”是利用 WordNet 网和 HowNet 网对词语进行语义分析并获得语义扩展的词语，用在对文本分析的图像检索系统中，并发明一种对返回结果进行优化排序的算法。通过语义扩展，可以很好认知用户的查询意图，不过这种方法的扩展词不对待查询的数据库进行分析，检索性能的通常会很有限；而且建立领域本体库即费力又费时。

2 情感实体抽取技术

情感对象就是情感表达作用的对象，通常为一个名词或名词性短语。通常情况下如果不知道情感对象，而仅仅进行情感倾向分析和研究是没有

价值的。情感对象的抽取作为情感分析和观点挖掘中一个非常重要同时也颇具挑战性的任务得到相关研究者的关注。尽管目前已经有许多情感表达和情感对象方面的研究，但是他们大多是针对产品评论信息或者新闻信息进行分析。

与传统的文本信息不同，微博由于系统字数的限制和网络行文的自由性，微博数据由于字数限制和行文自由性等原因，使之含有大量缩略的表达、错别字、特殊符号（如表情符号、链接等）等各类不同于传统规范的文字表达，这些无疑都提高了数据分析的难度。由于国内的情感分析和观点挖掘起步较晚以及中英文的差异性，另外相关技术的不成熟的限制，目前针对微博进行情感对象识别方面的研究还比较少。

目前已有的情感对象识别技术有北京航空航天大学申请的专利号为CN201210317183.0，发明名称为“基于词语依存关系的观点抽取方法”的专利。该方法采用基于词语依存关系链的匹配算法抽取评价对象，没有用到其他更多可用的辅助信息提高方法的准确度，其次该方法不一定适合于微博这种特殊的文本信息。

现有参考文献中常见的情感对象抽取主要针对的产品评论进行，由于有指定产品信息和领域限定，问题更加具体、清晰，因此主题相关文本的抽取工作往往都能达到比较好的效果。但是在其他主题无关的文本中效果并不佳，这主要在于这些文本中评论对象很杂，另外情感词也多样化。目前针对主题无关的微博进行情感对象识别技术很少，已有的方法大多是直接通过对微博进行句法依赖关系分析结合情感词典得到成对的<情感词，情感对象>关系，从而抽取情感对象。这种方法的识别效果不是很理想，存在

以下几个不足之处：（1）该抽取过程过分依赖于情感词典和特定几种句法依赖关系，一方面，由于基于词典的判断方法是有限的，且受到领域知识的影响很大，因此会存在很多的误判；另一方面，微博文字表达的特殊性，情感词和情感对象并不一定会局限于特定的几种依赖关系；（2）在微博中，常常一些情感词和其情感对象没有直接成对的出现在文本中，只有情感词表达情感倾向，而情感对象未显性的出现在句子中，那么该抽取过程不能够抽取一些没有直接出现在句子文本中的情感对象。

3 情感极性判别技术

目前已有情感分析系统及技术从分析的粒度上看主要集中于篇章级别以及句子级别的情感分析，而极少数的实体级别的情感分析技术将实体识别和情感分析分为两个独立的任务来进行。从分析的对象来看目前的系统及技术要针对新闻、微博等评论信息，关注于社会舆情的分析。

目前已有的篇章级别及句子级别情感分析技术主要有：西北工业大学的应用号为 CN200910219161.9、发明名称为“基于混合模型的 WEB 文本情感主题识别方法”的专利；中国科学院计算技术研究所的申请号为 CN200910083522.1、发明名称为“文本情感倾向性分析方法”的专利申请；中国科学院自动化研究所的申请号为 CN201210088366.X、发明名称为“一种面向微博短文本的情感分析方法”的专利申请；富士通株式会社的申请号为 CN201010157784.0、发明名称为“情感倾向性分析方法和装置”的专利申请。

上述情感分析技术主要包括训练和情感判断两个步骤，下面以西北工业大学的“基于混合模型的 WEB 文本情感主题识别方法”为例介绍其

在训练和情感判断的主要步骤，其余相关技术基本类似。该方法主要包括以下几个步骤：1、对训练集中的文本进行手工标注，估计出两类情感模型：“褒义”模型和“贬义”模型；同时根据不同主题文本的语言表达方式，分别估计各类主题语言模型；2、采用最大似然估计（MLE）方法对于步骤1建立的情感模型和主题模型分别进行参数估计；3、对于待处理的文本，计算其语言模型与两类情感模型的距离，从而对文本的情感倾向性以及主题进行判断。

目前的情感倾向性技术主要集中于篇章级别以及句子级别，基于机器学习的方法很普及，而基于情感落点的情感分析技术很少。

现有的基于情感词的情感分析技术主要存在以下三个方面的不足：A) 情感词组的提取没有考虑副词的修饰，但是一般情况下副词都会对形容词这类情感词产生程度限定作用。如果不加以考虑，容易造成情感强度偏差；B) 否定词的识别及处理问题，一般的方法是采取一种搜索的策略去查找否定词，很难确定否定的对象；C) 一些基于自动生成的情感词强度词典不可靠，因为情感词强度是情感词的基本属性，主要由其本意决定。

发明内容

本发明的目的在于克服现有情感实体搜索技术存在的上述不足，提出一种提升了情感极性判断的准确率的面向微博的情感实体搜索系统。

本发明通过以下技术方案实现：本发明面向微博的情感实体搜索系统，包括以下5个模块：

1) 用户接口，用于系统与用户的交互，用户可以通过该模块提交查询请求并获得反馈结果；

2) 查询扩展模块, 用于对微博语料数据进行词语关系挖掘, 并结合 WordNet 本体库建立加权词语关系图;

3) 查询处理模块, 用于将用户查询请求转换为索引库所能接受的查询关键词及查询语句, 并基于模块 2) 构建的词语关系图进行查询扩展;

4) 情感信息挖掘模块, 用于对微博语料库进行情感挖掘, 并生成情感实体及情感极性的判定规则;

5) 情感信息判定及索引建立模块, 用于对微博数据进行情感实体和情感极性的判定, 建立情感信息索引, 并进行存储;

6) 倒排索引建立模块, 用于对微博文本信息建立倒排索引, 并进行存储。

上述模块 1) 中采用以下步骤实现查询扩展:

11) 对微博语料库中的数据进行相关规则挖掘, 输出相关规则挖掘所得到的相关词集;

12) 结合 11) 所获得的频繁项及和 WordNet 本体库, 构建加权词语关系图。

上述步骤 11) 中采用 Eclat 算法挖掘微博语料库的频繁项集并生成相关词集, 并将相关词集和 WordNet 本体图通过映射或插入等形式形成加权词语关系图;

上述构建加权词语关系图时, 节点权重的计算方法为:

$$f(d) = \deg(d) = \deg^+(d) + \deg^-(d),$$

其中 $\deg(d)$ 、 $\deg^+(d)$ 、 $\deg^-(d)$ 分别表示结点的度、出度和入度; 边权重的计算方法为:

$$g(d_i, d_j) = \begin{cases} 1, & \text{若 } d_i, d_j \text{ 为原 WordNet 图的节点} \\ \text{lift}(d_i \rightarrow d_j), & \text{若 } d_i, d_j \text{ 为仅是规则词形成的节点} \\ \text{lift}(d_i \rightarrow d_j) + 1, & \text{若 } d_i, d_j \text{ 既是 WordNet 图节点也是规则词形成的节点} \end{cases}$$

。

上述模块 3) 中采用以下步骤实现查询处理:

31) 接收用户输入的查询词或语句;

32) 对用户的输入进行分词、去停用词和确定中心词的处理, 得到一个或多个中心词;

33) 将中心词在由本体和规则词构造的加权词语关系图库中选取适当

的扩展词,并对扩展词进行权重计算;

34) 然后选取权重大的前 p 个词语加入到查询词集合, 并将扩展词集合输入至查询接口。

上述步骤 33) 采用以下方法对扩展词进行权重计算:

假设原查询词为 $q = (q_1, q_2, \dots, q_m)$, 其中项 q_i 有 n_i 个最邻近词 $d_i = (d_{i1}, d_{i2}, \dots, d_{in_i})$, 则原查询项 q_i 与最邻近词项 d_{ij} 的相关度由计算方法为

$$W(q_i, d_{ij}) = \frac{g(q_i, d_{ij}) \times \log_2[f(d_{ij}) + 1]}{\sum_{j=1}^{n_i} \{g(q_i, d_{ij}) \times \log_2[f(d_{ij}) + 1]\}},$$

其中 $w(q_i, d_{ij})$ 为词 q_i 与词 d_{ij} 的相关度, $g(q_i, d_{ij})$ 为两词的权值, $f(d_{ij})$ 为词 d_{ij} 的度数, 所有最邻近词的权重计算方法为

$$W(d_k) = \sum_{d_{ij}=d_k} W(q_i, d_{ij}) / m。$$

上述模块 4) 中采用以下步骤实现情感实体的识别和判定:

- 41) 采集具有代表性的微博数据;
- 42) 对采集到的微博数据进行预处理, 包括对数据进行清洗、转化、分句、分词、词性标注以及句法解析等;
- 43) 对微博数据进行特征抽取, 将其表达成特征向量;
- 44) 训练情感实体识别模型, 获得模型参数;
- 45) 输出情感实体判定模型并存储。

上述步骤 43) 中采用以下方法实现特征抽取: 结合词语上下文, 设计包含全局特征在内的自定义词典, 根据自定义词典对微博数据进行特征抽取, 将微博数据转化为情感实体识别模型能够处理的输入数据格式。

上述步骤 44) 中采用以下方法实现情感实体识别模型: 在条件随机场 (CRF) 模型中引入全局特征节点, 建立结合全局特征的 GLCRF 模型, 并使用 L-BFGS 算法训练获得模型参数。

上述模块 5) 中采用以下步骤实现微博情感极性的判定:

- 51) 微博数据噪声去除及语义形式转化;

- 52) 分词, 词性标记及中文语法解析;
- 53) 结合情感词典提取情感词组;
- 54) 情感词组过滤;
- 55) 情感极性判定及结果输出。

上述步骤 53) 中采用 sentiPY 方法提取情感词组, 情感词组的形式统一表达为 *phrase:modifier*sentiment*, 即一个词组包括一个中心情感词, 同时可能附带多个修饰副词;

上述步骤 55) 中采用基于情感落点的混合决策算法对微博情感极性进行判定, 判定过程包含以下步骤

551) 判断句子中是否有概括词, 如无, 转步骤 552); 如有, 则以概括词之后的语句作为情感落点, 将情感落点极性作为微博情感极性输出;

552) 将微博句首及句尾作为情感落点, 比较句首、句尾情感极性, 若两者情感极性相互抵消, 则转 553); 否则, 将情感极性较强者优先作为微博情感极性进行输出;

553) 计算整条微博的情感词强度, 求和并平均, 将平均强度作为微博情感极性进行输出。

本发明针对微博情感实体搜索的查询扩展方案, 特征在于对微博语料数据进行词语关系挖掘, 结合 WordNet 本体库建立加权词语关系图, 并根据所构建的词语关系图进行查询扩展, 以更好的理解用户的查询意图; 本发明在查询扩展方面解决了语义本体与语料词语关系有效结合的问题, 能够更好的理解用户的查询用途, 进而将查询语句转化为更合适的查询扩展词; 在情感实体抽取及情感色彩分析方面, 解决了微博这类行文自由度较大的文本情感对象的抽取和情感极性的判断问题, 解决了的情感对象隐藏情况下的实体抽取问题, 优化了情感实体的抽取效果, 同时提升了情感极性判断的准确率。为网络舆情监控和产品舆情分析提供了一种优良的解决方案。本发明解决微博情感实体抽取、情感极性分析和情感实体搜索等困难问题,

为社交网络舆情分析和监控提供一种智能搜索产品。

附图说明

图 1 为本发明整体结构图；

图 2 为本发明的实施使用流程图；

图 3 为本发明的系统搭建构架图；

图 4 为本发明的情感极性分析方法的流程图；

图 5 为情感强度优化中基于相邻关系的图结构实例；

图 6 为情感落点算法流程图；

图 7 为微博情感对象抽取工作流程图；

图 8 为数据预处理流程图；

图 9 为情感对象模型训练实现原理图；

图 10 为 GLCRF 模型的图结构；

图 11 为 GLCRF 模型拓展多个全局节点后的模型图结构。

具体实施方式

以下结合附图对本发明的实施方式作进一步说明，但本发明的实施不限于此。

图 1 所示为本发明整体结构图。一种面向微博的情感实体搜索系统，包括：用户接口模块，用户可以通过该模块提交查询请求并获得反馈结果；查询扩展模块，实现对微博语料数据进行词语关系挖掘，并结合 WordNet 本体库建立加权词语关系图；查询处理模块，用于将用户查询请求转换为

索引库所能接受的查询关键词及查询语句，并基于查询扩展模块构建的词语关系图实现查询扩展；情感信息挖掘模块，用于对微博语料库进行情感挖掘，并生成情感实体及情感极性的判定规则；情感信息判定及索引建立模块，用于对微博数据进行情感实体和情感极性的判定，建立情感信息索引，并进行存储；倒排索引建立模块，用于对微博文本信息建立倒排索引，并进行存储。

图 2 示出了本发明查询处理模块的工作流程图。

参照图 2，该流程包括以下步骤：1、查询界面接收用户输入的查询词或语句；2、经过查询过程对用户的输入进行分词、去停用词和确定中心词的处理，得到一个或多个中心词，中心词可以是关键词也可以使修饰词等类型；3、将中心词在由本体和规则词构造的加权词语关系图库中选取适当的扩展词来源，选取的词语距离是 1，即是中心词的最邻近词；4、由于第三步所得的扩展词可能很多，因此为了衡量各个词语的重要性，对每个词语进行权重计算，然后选取权重大的前 p 个词语加入到查询词集合中；5、在第四步已经得到了所需要的扩展词，但是要引入一个机制可以让用户了解这些扩展词，并对词语进行操作，也就是对修改扩展的查询词集合，使得扩展词都符合用户查询意图；6、将扩展词集合返回查询入口，对富媒体数据库进行扩展检索；7、将检索的结果返回并显示给用户。

图 3 示出了本发明的查询处理和查询扩展模块的整合细节。

参照图 3，本发明的查询处理和查询扩展包括后台信息处理过程和检索过程两大部分，其中还可以分为微博信息抽取模块、建立索引模块、构建词语关系图模块、用户检索模块以及管理员操作和用户操作模块五大子

模块。

微博信息抽取模块的过程包括组织好初始的微博数据、对其进行适当的清洗、分句、分词和语法分析。建立索引模块主要是对微博数据集建立一个索引供快速检索。我们采用 Lucene 来建立倒排索引。Lucene 是一个开源的全文检索引擎的架构，提供了完整的查询引擎和索引引擎，支持布尔操作、模糊查询、分组查询等等操作。用其建立好倒排索引并保存。

构建词语关系图库模块是本文的核心部分，也是创新的部分。该部分分为分词过程、Eclat 相关规则挖掘过程、相关规则词生成过程以及结合 WordNet 生成加权词语关系图过程。分词过程就是将文本的文字资源分成一个个词语。我们采用对中文分词精确率较高的 ICTCLAS 软件进行分词，这是中科院研发的专门针对中文分词的系统。我们先逐个地对数据集的文档进行分词，然后再将各种类的文档合在一起形成一个文档集，以供相关规则挖掘使用。在相关规则的挖掘过程中，我们采用挖掘效率较高的 Eclat 挖掘算法，这是一个深度优先的算法，对大文档可以分块的挖掘相关词最后再合并起来。本发明使用支持度-兴趣度的相关规则框架，该框架采用两条评判公式：

(1)、支持度公式：

$$supp(X \rightarrow Y) = \frac{|X \cup Y|}{|D|}$$

(2)、兴趣度公式：

$$lift(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)}$$

其中 $|X \cup Y|$ 是同时包含 X 和 Y 的事务数, $|D|$ 是数据库的事务总数; $supp(X \cup Y)$ 是数据库中事务同时包含 x 和 y 的百分比, $supp(X)$ 、 $supp(Y)$ 分别表示事务只包含 x 和只包含 y 的百分比。

在挖掘过程中根据不同的文档集设定了不同支持度阈值,而挖掘出的频繁项集只有在兴趣度大于1时才生产相关规则项。因为本发明认为只要当两个词的兴趣度大于1时他们才是正相关的。在挖掘过程还添加了合成词的概念:当两个词的兴趣度值大于4时,将这个规则项的前后两个词合并生成组合词,这个词分别与规则词的前件和后件形成一个新规则,新规则的兴趣度值与原规则的相同,这样合成词也可作为扩展词被选取。在相关词语挖掘出来后将生产相关规则词并保存,保存的格式是“ $x \ y$ ”的形式。此时即完成了相关规则词的挖掘及分析。

剩下的一步是将这些规则词和 WordNet 本体库结合成一个加权词语关系图。WordNet 是基于词汇的语义网络。WordNet 不但将词汇组织成概念,还定义了概念、词汇间的多种语义关系(如同位词、上/下位词、反义词、整体-部分词、蕴涵等等),词与词的关系形成一个有向图(如图3的示例)。此过程我们考虑将规则词项按一定的顺序映射或添加到 WordNet 本体库中,我们设定加权词语关系图的构造原则是:在两个规则词的结点间添加一条由前件指向后件的有向边。其中规则词的添加完全自动化,分为两种情况:第一,若原 WordNet 本体图中存在这个词,则只需将词映射到图即可,然后更新结点数据;第二,若原 WordNet 本体图中不存在这个词,则先添加词语,再添加边以及更新数据。所有结点数据在图完成后逐一统计。最终形成的关系图可以用一个四重组表示: $G = \langle V, E, f, g \rangle$, 其中 v 是结点

集合, E 是边的集合, f 是从 V 到非负实数集合的函数, 设为结点的度数; g 是从 E 到非负实数集合的函数, 设为两个结点边的值。设 $d, d_i, d_j \in V$, $\deg(d)$ 表示结点 d 的度(即该结点的出度和入度之和), $\text{lift}(d_i \rightarrow d_j)$ 表示结点词语 d_i 、 d_j 的兴趣度值, 那么有:

$$(1)、f(d) = \deg(d)$$

$$(2)、g(d_i, d_j) = \begin{cases} 1, & \text{若 } d_i, d_j \text{ 为原 WordNet 图的节点} \\ \text{lift}(d_i \rightarrow d_j), & \text{若 } d_i, d_j \text{ 为仅是规则词形成的节点} \\ \text{lift}(d_i \rightarrow d_j) + 1, & \text{若 } d_i, d_j \text{ 既是 WordNet 图节点也是规则词形成的节点.} \end{cases}$$

加权词语关系图(如图 4 的示例)中, 词在整个图中的重要程度由该词所在结点的度衡量, 即结点的出度和入度之和(图 4 中结点旁的整数值);边的值是权值, 其中原 WordNet 图的本体词间的权值设为 1(图 4 中蓝色的边), 由规则插入的词间的权值设为两词的兴趣度值(图 4 中蓝色的边), 若两词既是 WordNet 关系词又是规则词, 则权值为兴趣度值加 1。图 4 中黑色边所指的词为合成词(如“知识产权”), 其与两个规则词的权值是相同的。此时即完成了加权词语关系图的构建。

用户检索模块包括查询输入、查询分析过程、匹配扩展词语过程、生成扩展查询词集合过程、检索索引过程以及结果处理并显示给用户的过程。查询输入就是在查询界面接收用户输入的查询词或语句; 查询分析即是用户的输入进行分词、去停用词和确定中心词的处理, 得到一个或多个中心词; 匹配扩展词语过程是将上一步骤的中心词输入到加权词语关系图库中选取适当的扩展词来源, 即从这个图中选取距原查询词最近的词(即距离为 1 的词)作为候选扩展词。生成扩展查询词集合过程是根据各个词与原查询词的相关度, 计算词的权重后选取前 p 个作为最终扩展词。本发明创建了计算

各词语权重的公式,根据加权词语关系图的结构可知:如果两个结点的权值越大,表示这两个结点的相关度也越大;而如果结点的度越大,表明该结点的重要性也越大.

假设原查询词为 $q=(q_1,q_2,\cdots,q_m)$, 其中项 q_i 有 n_i 个最邻近词 $d_i=(d_{i1},d_{i2},\cdots,q_{in_i})$,

则原查询项 q_i 与最邻近词项 d_{ij} 的相关度由计算方法为

$$W(q_i,d_{ij})=\frac{g(q_i,d_{ij})\times\log_2[f(d_{ij})+1]}{\sum_{j=1}^{n_i}\{g(q_i,d_{ij})\times\log_2[f(d_{ij})+1]\}},$$

其中 $w(q_i,d_{ij})$ 为词 q_i 与词 d_{ij} 的相关度, $g(q_i,d_{ij})$ 为两词的权值, $f(d_{ij})$ 为词 d_{ij} 的度数, 所有最邻近词的权重计算方法为

$$W(d_k)=\sum_{d_{ij}=d_k}W(q_i,d_{ij})/m。$$

其中 $w(d_k)$ 为词 d_k 的权重, m 表示原查询词的个数。在算出各个候选扩展词的权重后,将权重按降序排列,并选取前 p 个词加入到原查询中,构成扩展词集合,其中原查询项的权重都为 1。

由上一步以及得到扩展词集合,如以下的形式:

$$Q=(q_1,q_2,\dots,q_m,d_1,d_2,\dots,d_p) \quad (4)$$

检索过程是指将扩展词集合返回查询入口返回查询入口,对富媒体数据库进行扩展检索。结果处理并显示的过程是指将排好序的检索的结果返回并显示给用户。

图 4 为本发明提出的情感极性分析方法的流程图。

参照图 4,该方法包括以下步骤:

(1) 评论语料的噪声去除及语义形式转化：

评论语料的噪声去除主要是除去干扰子句如虚拟语气。这些干扰的句子非真实客观的评价，会干扰后阶段的分析。替换表情符号为相应的文字，从而将语义形式转化为友好处理的形式。

(2) 自然语言处理：主要是利用 Stanford NLP 软件对评论语料进行分词，词性标记及中文语法解析。

(3) 结合情感词典提取情感词组：

因为情感词在评论语料中的 POS tagger label 主要集中在少数几个 label 上面，我们就结合这些词性标签和情感词典提取情感词组。采用我们开发的 sentiPY 方法提取情感词组，在本系统情感词组的形式统一为：

*phrase: modifier * sentiment*

，即一个词组包括一个中心情感词，可能附带多个修饰副词。

(4) 情感词组过滤：对第 3 步中提取的粗粒度情感词组进行过滤，使得情感词组的形式更纯，从而可以提升最终的极性分类的准确度。

(5) 情感分析并将结果输出

我们设计了一个基于情感落点的混合决策算法，该算法可以有效的对不同领域的评论语料进行分析。

图 5 为情感强度优化中基于相邻关系的图结构实例。参照图 5，把评论语料中的情感词看作是图中的节点，基于传播的算法可以计算上下文的情感强度。基于情感词典，提取情感词相邻的关系并通过 NGD 计算两情感词节点的权重，从而形成一个有向图。图三为一条评论的图结构。

图 6 为情感落点算法流程图。参照图 4，在该步骤中，我们的目标是找到一条评论的情感落点。所谓的情感落点就是在一条评论中作者主要想表达的情感部分。我们主要依据概括性的词汇（如“总体”）、比较开头结尾处的情感强度及句子中的最强情感词组，从而找到一条评论的情感落点。

图 7 示出了本发明针对微博情感实体抽取的工作流程图。

参照图 1，本发明的情感实体抽取包括微博数据采集、数据预处理、特征提取、词典加载、标记与修正、模型训练和情感对象抽取等步骤。微

博数据采集从互联网爬取的微博数据将以文件的形式保存起来，模型训练得到的情感对象抽取模型也会被保存起来用于对象抽取，情感对象抽取得到的结果以将文件的形式保存下来，以使用户查看和修正预测结果。

微博数据采集，用于从互联网上的微博系统（如新浪微博、twitter 和腾讯微博等）爬取微博数据，并将采集下来的微博原始数据按照一定的组织方式以文件的形式保存下来，为系统的后期处理提供数据支持。

数据预处理，用于对原始的微博数据进行一些预先处理，便于后期进行特征提取。该模块包括数据清洗、数据转化、分句、分词、词性标注和句法解析。详情如图 2 所示。

词典加载，用于加载数据预处理和特征提取步骤所需要的相关词典，这项词典包括情感词典、停用词词典、常见网络用语词典等词典数据。

特征抽取，借助词典加载模块加载的词典数据对与处理后的数据进行预先定义特征的抽取，将文本向量化，转化为对象抽取模块能够处理的格式。

情感对象模型训练，用于本系统核心的情感对象抽取模型进行训练。从标记和修正模块获取转化为要求格式的训练数据，使用 L-BFGS 算法对根据训练数据构建的 CRF 模型进行训练。本发明使用的 CRF 模型是在 Linear CRF（线性条件随机场）模型的基础上演变而来，是 CRF（条件随机场）模型第一次在情感对象识别领域进行应用。通过在传统的 CRF 模型中添加全局变量，从而达到能够识别出情感对象不显性出现在标记序列中的情况。

情感对象抽取，用于从微博数据中抽取出情感对象，该步骤主要利用模型训练模块训练出的模型来进行预测从而达到抽取对象的目的。

标记和修正，本发明中用到的 CRF 模型为一个有监督统计学习方法，因此需要对数据进行标注。同时引入反馈机制对错误分析信息进行学习。现有方法对于误分结果一般不作处理，但这些反馈信息包含了大量有用信息，如何能够充分利用这些信息成了系统实现自我学习的关键。反馈机制的引入使得模型能够对错误分析的结果进行再次学习，使得系统越用越准。

图 8 示出了本发明数据预处理步骤的实现原理图，数据预处理步骤包括以下步骤：

(1) 数据清洗处理步骤，从数据采集模块收集的原始微博数据中读取数据，进行数据预处理中的数据清洗过程，过滤掉一些空的、无效的脏微博数据。

(2) 数据转化处理步骤，该步骤处理从(1)步骤处理后传过来的数据，对微博数据中的一些内容进行转化处理，便于(3)(4)(5)(6)步骤相关处理，常见有以下几种情况：(a) 微博中常常含有一些对工作无效的信息，则需要剔除掉；(b) 一些对我们工作来说无用的链接（如图片链接和网页链接等）和特殊字符串需要剔除掉；(c) 在微博中常常包含有带“#”符号的话题和带“@”符号的联系人也进行了处理，我们把微博头和尾出现的话题和联系人直接删除，在微博句子中的则只删除“#”和“@”符号；(d) 微博中常常包含有一些表情符号，这些符号是带有强烈的情感倾向的，也是对我们的工作有帮助的信息，但是这些符号会影响分词、词性标注（POS 标注）和句法解析的精度，因此在此过程中需要提取出来；(e) 需要对微博中一些网络用语进行转换，例如，把网络表达方式的“V5”转成规范表达的“威武”等，这同样有助于提高分词、词性标注（POS 标注）和句法解析的精度。

(3) 微博文本分句处理步骤，本发明的情感对象识别方法的条件随机场模型是构建在句子级别的序列标记，进行信息抽取，然而一条微博可能包含有一个以上的句子，因此需要对之进行分句处理。在分句处理过程中主要是根据标点符号进行分句。但是由于微博的特殊性，仅仅根据标点符号进行分句是不够的。在微博中很多人为了方便，习惯用空格或者特殊的符号（如“~”等）进行分句，因此在此过程中还针对这些情况进行了对应分句处理。

(4) 句子分词处理步骤，本发明的情感对象识别方法的条件随机场模型是对句子级别的序列中每个词进行标记，因此需要进行分词处理。分词过程用到的是一些常用网络用语词汇词典（如“抓狂”、“围观”等）用于提高分词的准确度。

(5) 句子中词的词性标注步骤，此步骤对分词后的每个词进行词性标注，为本发明的特征提取模型进行特征提取时提供词的词性相关特征。

(6) 句法解析步骤，此步骤利用句法解析工具解析出句子中词之间的

句法依赖关系，目的为本发明的特征提取模型进行特征提取时提供词的依赖相关特征。

图 9 为本发明情感对象识别模型训练步骤的实现原理图。参照图 9，在该步骤中，已标注的训练数据集来源于数据采集模块从互联网中爬取的微博数据，并进行数据预处理模块处理。由于本发明中采用的条件随机场（CRF）模型进行情感对象抽取，而 CRF 模型为一种有监督学习方法，因此在训练过程中的训练数据集还需要进行人工标注数据集。训练模型过程中，首先需要利用词典加载模块加载用户词典，包括情感词词典和停用词词典；下一步就是利用特征提取模块结合上一部加载的词典对训练数据集进行特征提取并规范化数据；最后一步是利用模型训练模块对上步规范化的数据进行模型参数训练，使用 L-BFGS 算法训练学习得到模型的参数。

在本发明中用到的条件随机场模型如图 10 所示的形式，把情感对象识别过程看成是一个序列标记问题。该模型的第一层的 X 表示输入的微博句子， x_i 表示句子中第 i 个位置的词，第二层的 y_1 和第三层的 g_1 、 g_2 输出结果状态，这些状态的标签的可能取值为： $L = \{ 'N-B', 'N-I', 'P-B', 'P-I', 'O' \}$ 这五个标签，它表示序列标记过程中序列每个位置标记标签取值空间，其中 $N-B$ 标签表示负向情感对象的开始位置标签， $N-I$ 标签表示负向情感对象的后继标签（即其前一个标签必须为 $N-B$ 或者 $N-I$ ）， $P-B$ 标签表示正向情感对象的开始位置标签， $P-I$ 标签表示正向情感对象的后继标签（同理前一个标签必须为 $P-B$ 或者 $P-I$ ）， O 标签表示其他所有标签，即有 $y_i \in L$ 。例如序列为 {“手机”，“屏幕”，“非常”，“清晰”}，“手机屏幕”为一正向的情感对象，对之进行标记的结果为 {“ $P-B$ ”，“ $P-I$ ”，“ O ”，“ O ”}。

模型中用两个全局节点 g_1 和 g_2 表示两个独立的单一情感对象，因此取值只能为 {“ $N-B$ ”，“ $P-B$ ”，“ O ”} 这三个标签，要么为正向情感对象即为 $P-B$ 标签，要么为负向情感对象即为 $N-B$ 标签，要么不是情感对象即为 O 标签，而不可能为情感对象的后继标签 $N-I$ 和 $P-I$ 。

为了提高情感对象识别的灵活性和可拓展性，本发明采用的条件随机场模型不局限于图 9 所示的图结果，表示非显性也不局限于两个隐藏的节点 g_1 和 g_2 ，可以拓展到如图 11 所示的 $g_1 \cdots g_n$ （ $n \geq 1$ ）。

以上所述的具体实施例，对本发明的目的、技术方案和有益效果进行了进

一步详细说明，所应理解的是，以上所述仅为本发明的具体实施例而已，并不用于限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

权 利 要 求 书

1、一种面向微博的情感实体搜索系统，其特征在于包括以下 5 个模块：

1) 用户接口，用于系统与用户的交互，用户可以通过该模块提交查询请求并获得反馈结果；

2) 查询扩展模块，用于对微博语料数据进行词语关系挖掘，并结合 WordNet 本体库建立加权词语关系图；

3) 查询处理模块，用于将用户查询请求转换为索引库所能接受的查询关键词及查询语句，并基于模块 2) 构建的词语关系图进行查询扩展；

4) 情感信息挖掘模块，用于对微博语料库进行情感挖掘，并生成情感实体及情感极性的判定规则；

5) 情感信息判定及索引建立模块，用于对微博数据进行情感实体和情感极性的判定，建立情感信息索引，并进行存储；

6) 倒排索引建立模块，用于对微博文本信息建立倒排索引，并进行存储。

2、根据权利要求 1 所述的面向微博的情感实体搜索系统，其特征在于上述模块 1) 中采用以下步骤实现查询扩展：

11) 对微博语料库中的数据进行相关规则挖掘，输出相关规则挖掘所得到的相关词集；

12) 结合 11) 所获得的频繁项及和 WordNet 本体库，构建加权词语关系图。

3、根据权利要求 1 所述的面向微博的情感实体搜索系统，其特征在于上述步骤 11) 中采用 Eclat 算法挖掘微博语料库的频繁项集并生成相关词集，并将相关词集和 WordNet 本体图通过映射或插入等形式形成加权词语关系图；

上述构建加权词语关系图时，节点权重的计算方法为：

$$f(d) = deg(d) = deg^+(d) + deg^-(d),$$

其中 $\deg(d)$ 、 $\deg^+(d)$ 、 $\deg^-(d)$ 分别表示结点的度、出度和入度；边权重的计算方法为：

$$g(d_i, d_j) = \begin{cases} 1, & \text{若 } d_i, d_j \text{ 为原 WordNet 图的节点} \\ \text{lift}(d_i \rightarrow d_j), & \text{若 } d_i, d_j \text{ 为仅是规则词形成的节点} \\ \text{lift}(d_i \rightarrow d_j) + 1, & \text{若 } d_i, d_j \text{ 既是 WordNet 图节点也是规则词形成的节点.} \end{cases}$$

其中 $\text{lift}(d_i \rightarrow d_j)$ 是根据 Eclat 算法所得的 d_i, d_j 的相关度。

4、根据权利要求 1 所述的面向微博的情感实体搜索系统，其特征在于上述模块 3) 中采用以下步骤实现查询处理：

- 31) 接收用户输入的查询词或语句；
- 32) 对用户的输入进行分词、去停用词和确定中心词的处理，得到一个或多个中心词；
- 33) 将中心词在由本体和规则词构造的加权词语关系图库中选取适当的扩展词，并对扩展词进行权重计算；
- 34) 然后选取权重大的前 p 个词语加入到查询词集合，并将扩展词集合输入至查询接口。

5、根据权利要求 4 所述的面向微博的情感实体搜索系统，其特征在于上述步骤 33) 采用以下方法对扩展词进行权重计算：

假设原查询词为 $q = (q_1, q_2, \dots, q_m)$ ，其中项 q_i 有 n_i 个最邻近 $d_i = (d_{i1}, d_{i2}, \dots, d_{in_i})$ ，则原查询项 q_i 与最邻近词项 d_{ij} 的相关度由计算方法为

$$W(q_i, d_{ij}) = \frac{g(q_i, d_{ij}) \times \log_2[f(d_{ij}) + 1]}{\sum_{j=1}^{n_i} \{g(q_i, d_{ij}) \times \log_2[f(d_{ij}) + 1]\}},$$

其中 $W(q_i, d_{ij})$ 为词 q_i 与词 d_{ij} 的相关度， $g(q_i, d_{ij})$ 为两词的权值， $f(d_{ij})$ 为词 d_{ij} 的度数，所有最邻近词的权重计算方法为

$$W(d_k) = \sum_{d_{ij}=d_k} W(q_i, d_{ij}) / m。$$

6、根据权利要求 1 所述的面向微博的情感实体搜索系统，其特征在于上述模块 4) 中采用以下步骤实现情感实体的识别和判定：

- 41) 采集具有代表性的微博数据；
- 42) 对采集到的微博数据进行预处理，包括对数据进行清洗、转化、分句、分词、词性标注以及句法解析等；
- 43) 对微博数据进行特征抽取，将其表达成特征向量；
- 44) 训练情感实体识别模型，获得模型参数；
- 45) 输出情感实体判定模型并存储。

7、根据权利要求 6 所述的面向微博的情感实体搜索系统，其特征在于上述步骤 43) 中采用以下方法实现特征抽取：结合词语上下文，设计包含全局特征在内的自定义词典，根据自定义词典对微博数据进行特征抽取，将微博数据转化为情感实体识别模型能够处理的输入数据格式。

8、根据权利要求 6 所述的面向微博的情感实体搜索系统，其特征在于上述步骤 44) 中采用以下方法实现情感实体识别模型：在条件随机场 (CRF) 模型中引入全局特征节点，建立结合全局特征的 GLCRF 模型 (全局条件随机场模型)，并使用 L-BFGS 算法训练获得模型参数。

9、根据权利要求 1 所述的面向微博的情感实体搜索系统，其特征在于上述模块 5) 中采用以下步骤实现微博情感极性的判定：

- 51) 微博数据噪声去除及语义形式转化；
- 52) 分词，词性标记及中文语法解析；
- 53) 结合情感词典提取情感词组；
- 54) 情感词组过滤；
- 55) 情感极性判定及结果输出。

10、根据权利要求 9 所述的面向微博的情感实体搜索系统，其特征在于上述步骤 53) 中采用 sentiPY 方法提取情感词组，情感词组的形式统一表达为 *phrase:modifier*sentiment*，即一个词组包括一个中心情感词 (sentiment)，同时可能附带多个修饰副词 (modifier)；

上述步骤 55) 中采用基于情感落点的混合决策算法对微博情感极

性进行判定，判定过程包含以下步骤

551) 判断句子中是否有概括词，如无，转步骤 552)；如有，则以概括词之后的语句作为情感落点，将情感落点极性作为微博情感极性输出；

552) 将微博句首及句尾作为情感落点，比较句首、句尾情感极性，若两者情感极性相互抵消，则转 553)；否则，将情感极性较强者作为微博情感极性进行输出；

553) 计算整条微博的情感词强度，求和并平均，将平均强度作为微博情感极性进行输出。

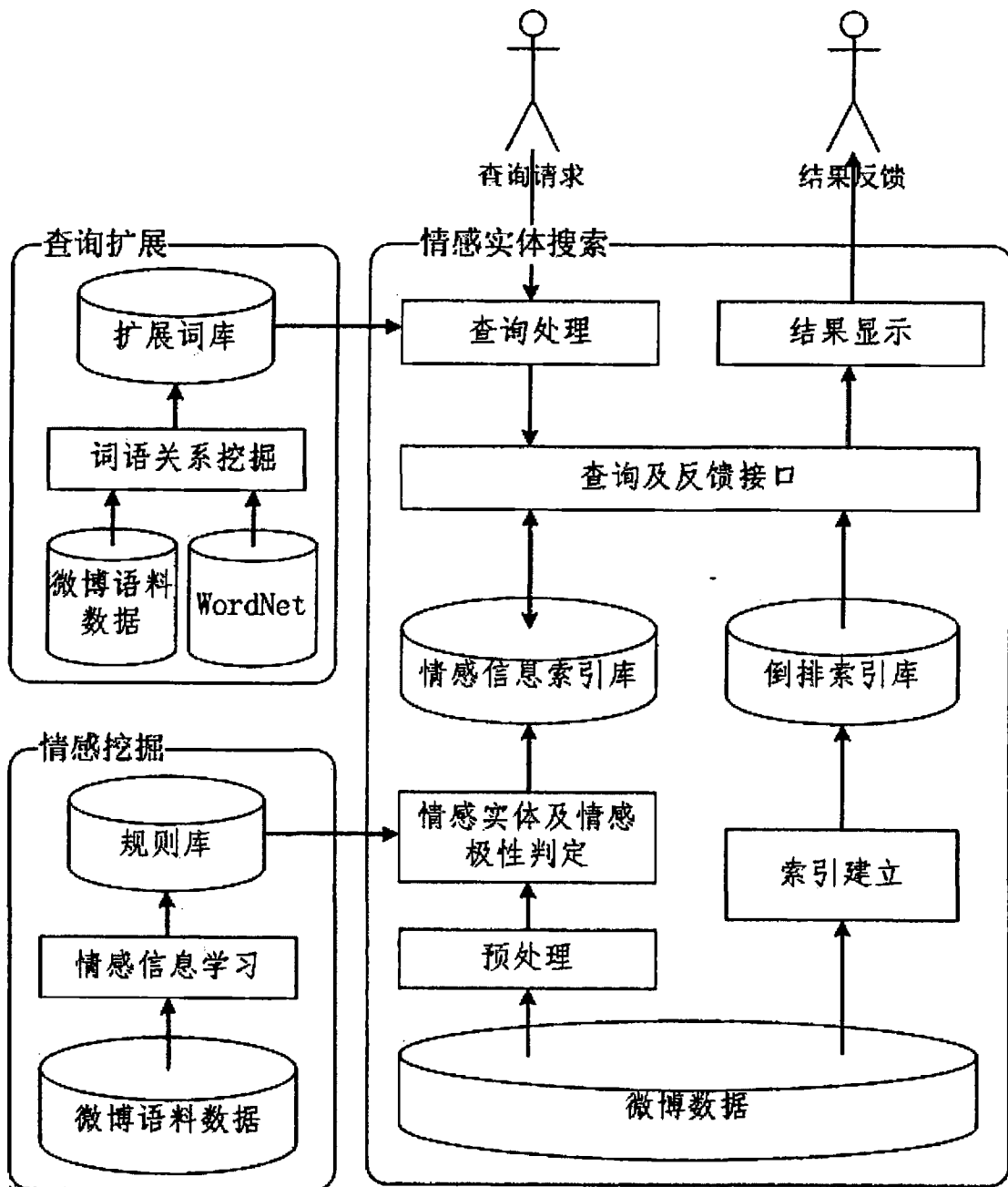


图 1

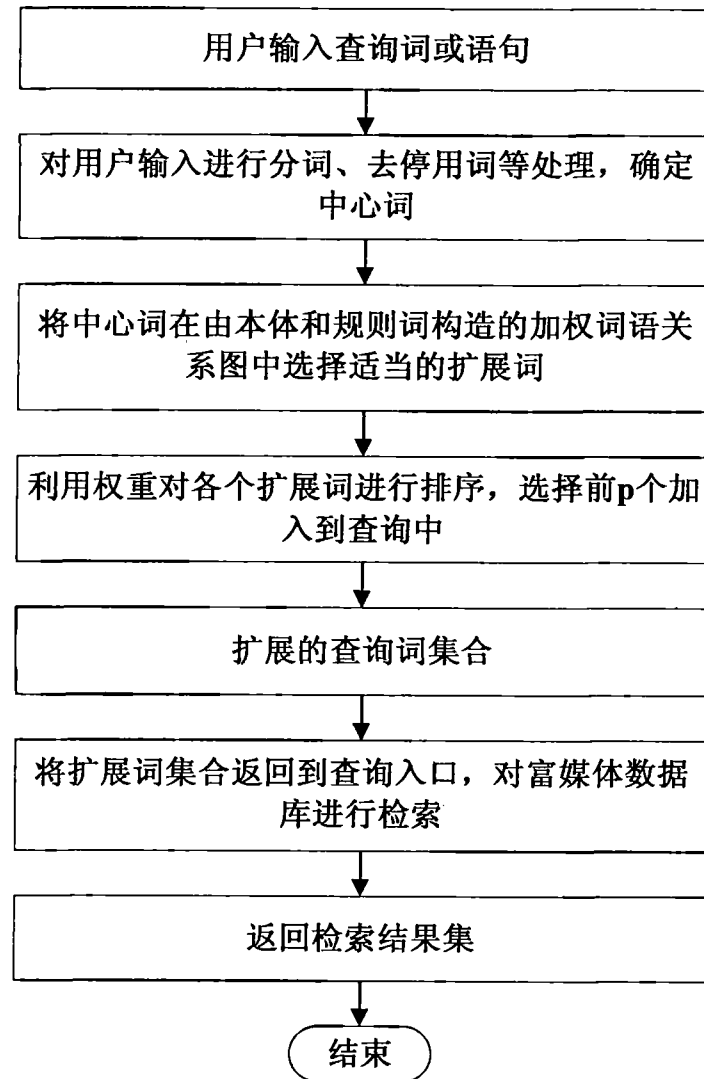


图 2

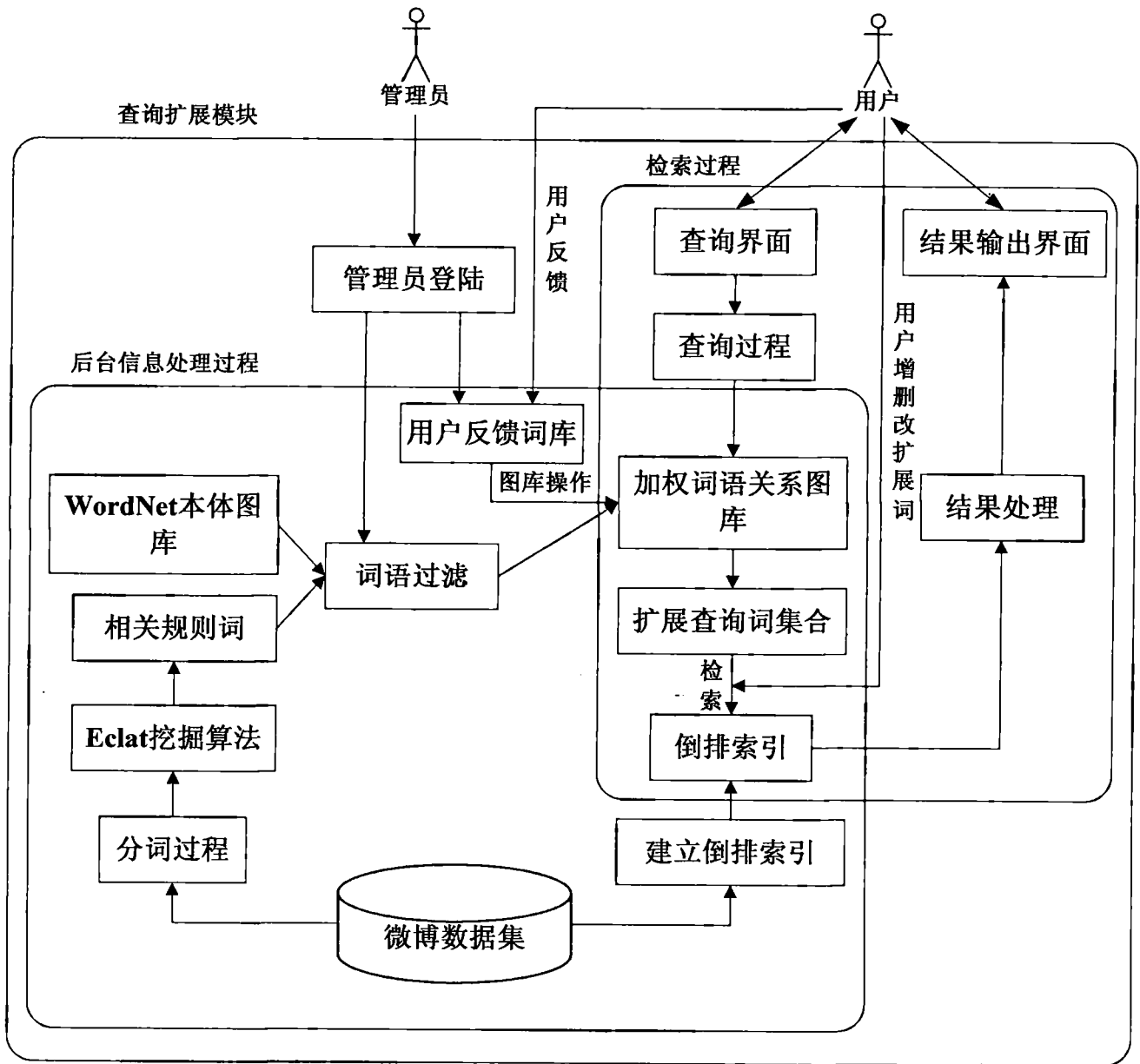


图 3

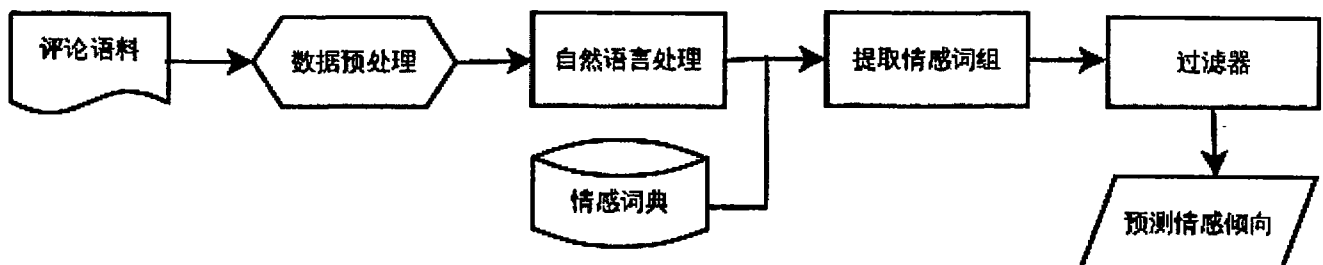


图 4

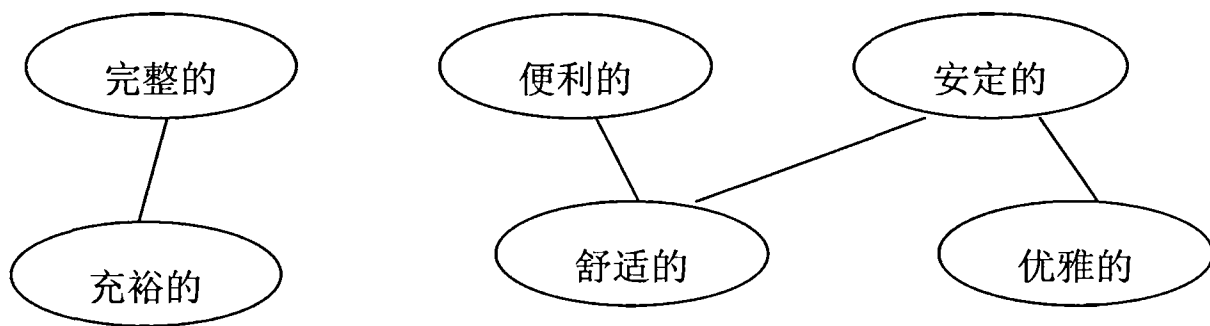


图 5

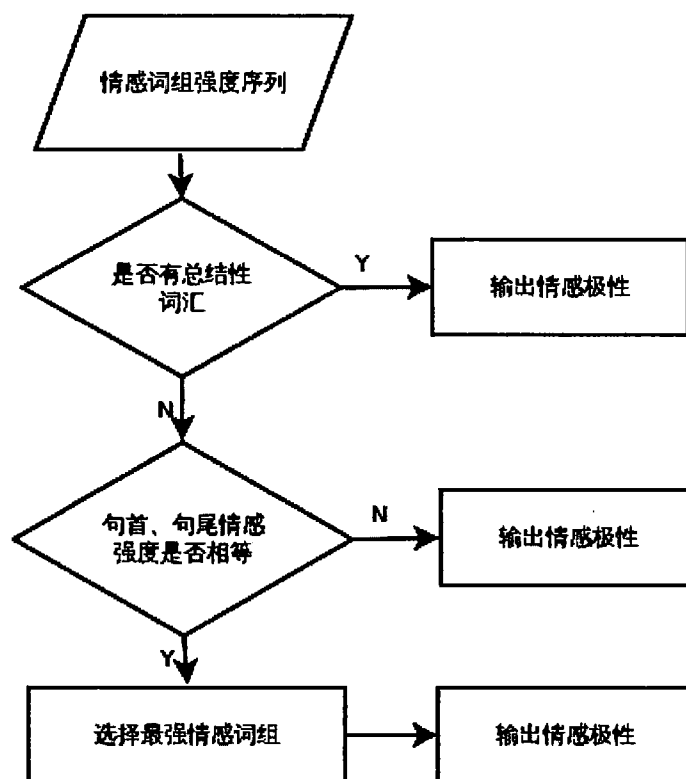


图 6

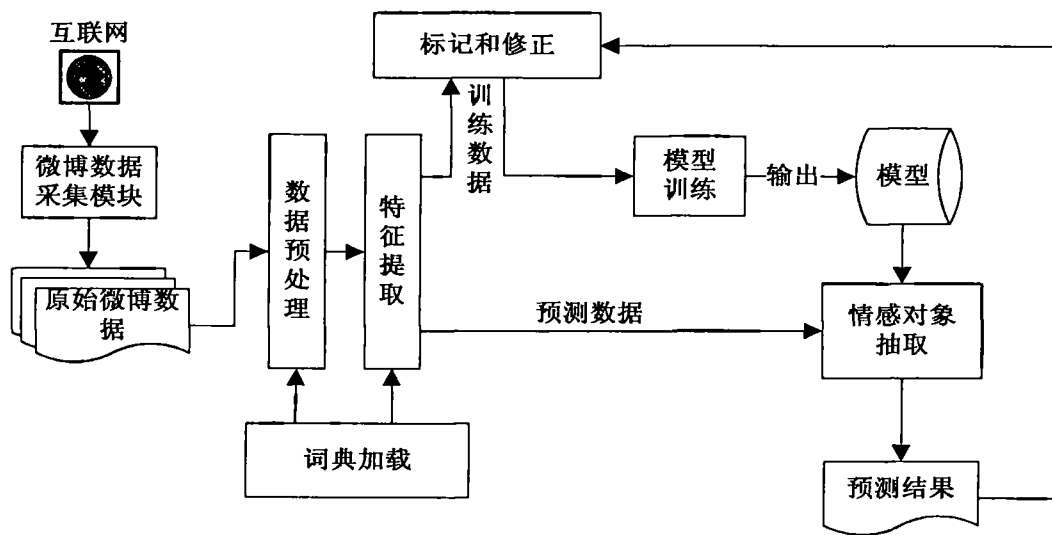


图 7

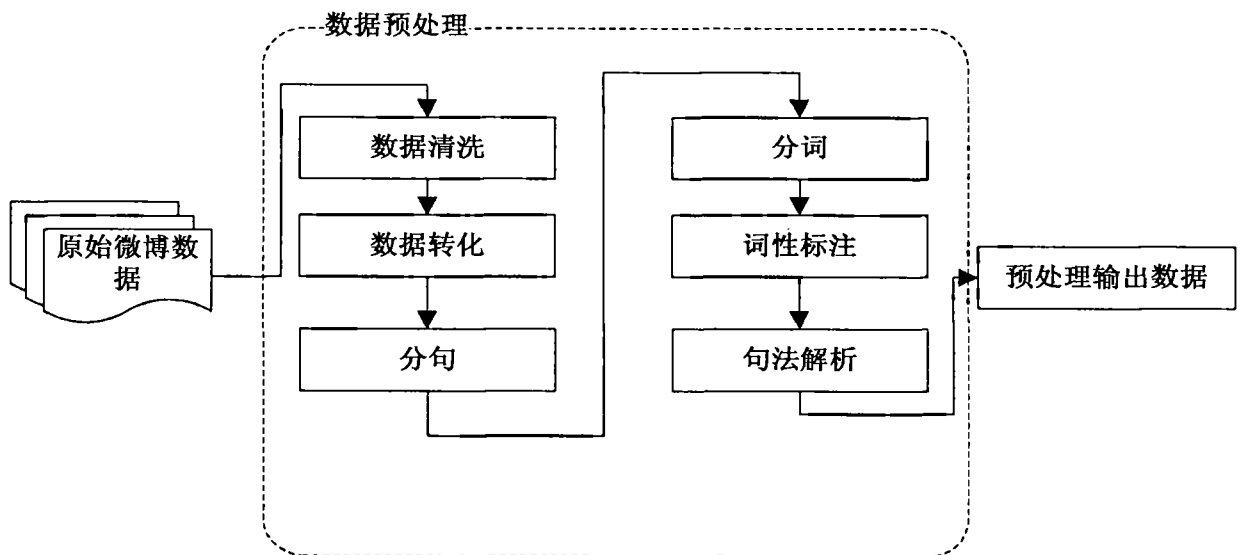


图 8

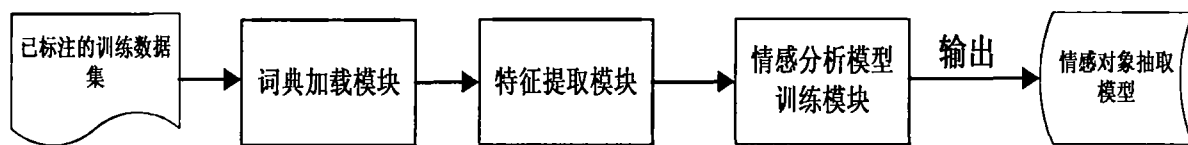


图 9

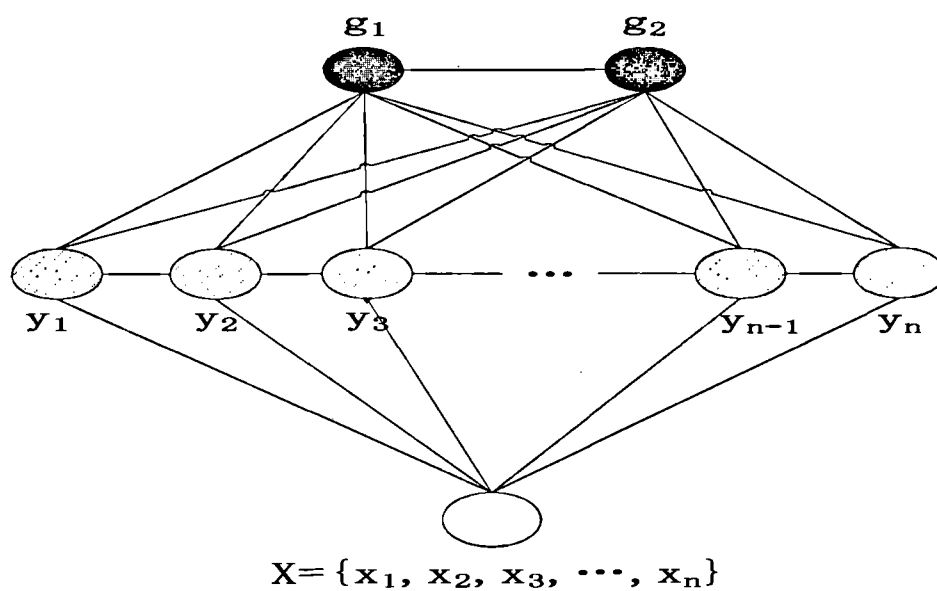


图 10

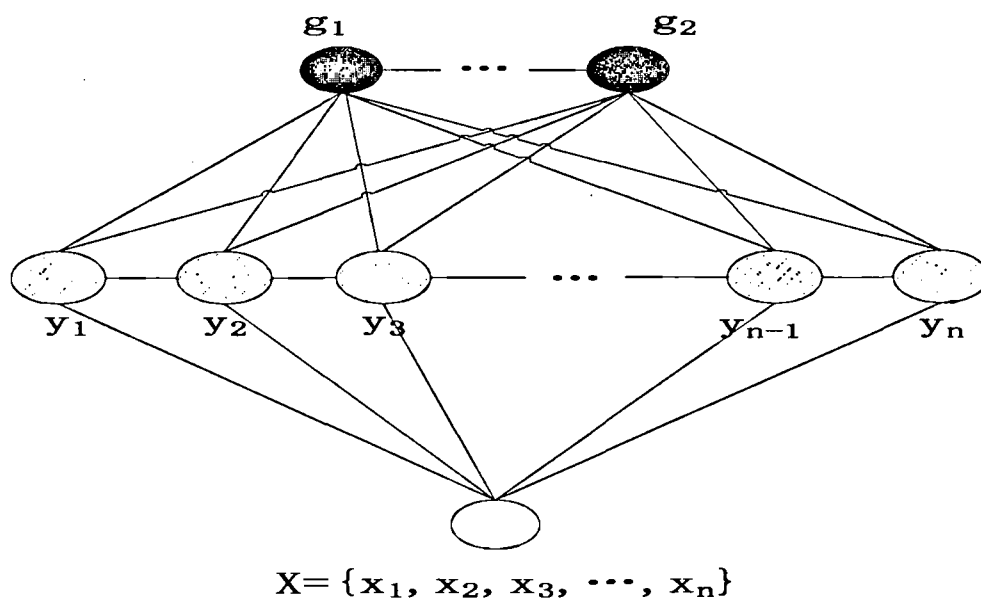


图 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2013/088772

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 17/-; G06Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

VEN, DWPI, CPRS, CNKI, PATENTICS: microblog, extend, ontology base, corpus, semantic, eclat, wordnet, blog, emotion, search, keyword, character, extent, query, model, relationship, relation, index

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 102279890 A (SOOCHOW UNIVERSITY), 14 December 2011 (14.12.2011), the whole document	1-10
A	CN 102073692 A (BEIJING RESEARCH CENTER FOR INFORMATION TECHNOLOGY IN AGRICULTURE), 25 May 2011 (25.05.2011), the whole document	1-10
A	WO 2011112319 A2 (YAHOO! INC.), 15 September 2011 (15.09.2011), the whole document	1-10

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“E” earlier application or patent but published on or after the international filing date	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“&” document member of the same patent family
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 27 June 2014 (27.06.2014)	Date of mailing of the international search report 07 July 2014 (07.07.2014)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer TIAN, Jing Telephone No.: (86-10) 62089417

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2013/088772

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102279890 A	14 December 2011	None	
CN 102073692 A	25 May 2011	None	
WO 2011112319 A2	15 September 2011	WO 2011112319 A3	17 November 2011
		US 2011225043 A1	15 September 2011
		TW 201203154 A	16 January 2012

国际检索报告

国际申请号

PCT/CN2013/088772

A. 主题的分类

G06F 17/30 (2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F17/-; G06Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

VEN, DWPI, CPRS, CNKI, PATENTICS: 微博, 情感, 搜索, 关键词, 关键字, 扩展, 查询, 模型, 关系, 本体库, 语料, 语义, 索引, eclat, wordnet, blog, emotion, search, keyword, character, extent, query, model, relationship, relation, index

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 102279890 A (苏州大学) 2011年 12月 14日 (2011 - 12 - 14) 全文	1-10
A	CN 102073692 A (北京农业信息技术研究中心) 2011年 5月 25日 (2011 - 05 - 25) 全文	1-10
A	WO 2011112319 A2 (YAHOO! INC) 2011年 9月 15日 (2011 - 09 - 15) 全文	1-10

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2014年 6月 27日

国际检索报告邮寄日期

2014年 7月 07日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局(ISA/CN)
北京市海淀区蓟门桥西土城路6号
100088 中国

授权官员

田竟

传真号 (86-10)62019451

电话号码 (86-10)62089417

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2013/088772

检索报告引用的专利文件		公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	102279890 A	2011年 12月 14日	无	
CN	102073692 A	2011年 5月 25日	无	
WO	2011112319 A2	2011年 9月 15日	WO 2011112319 A3	2011年 11月 17日
			US 2011225043 A1	2011年 9月 15日
			TW 201203154 A	2012年 1月 16日

表 PCT/ISA/210 (同族专利附件) (2009年7月)