(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2012 年 11 月 22 日 (22.11.2012)

WIPO | PCT

(10) 国际公布号
WO 2012/155714 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2012/073656

(22) 国际申请日: 2012 年 4 月 9 日 (09.04.2012)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201110124378.9 2011 年 5 月 13 日 (13.05.2011) CN

(71) 申请人 (对除美国外的所有指定国): **北京百度网讯科技有限公司 (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.)** [CN/CN]; 中国北京市海淀区上地十街 10 号百度大厦 2 层, Beijing 100085 (CN)。

(72) 发明人; 及

(75) 发明人/申请人 (仅对美国): **赵世奇 (ZHAO, Shiqi)** [CN/CN]; 中国北京市海淀区上地十街 10 号百度大厦 2 层, Beijing 100085 (CN)。 **柴春光 (CHAI, Chunguang)** [CN/CN]; 中国北京市海淀区上地十街 10 号百度大厦 2 层, Beijing 100085 (CN)。 **吴华 (WU, Hua)** [CN/CN]; 中国北京市海淀区上地十街 10 号百度大厦 2 层, Beijing 100085 (CN)。

(74) 代理人: **深圳市威世博知识产权代理事务所 (普通合伙) (CHINA WISPRO INTELLECTUAL PROPERTY LLP.)**; 中国广东省深圳市南山区高新区粤兴三道 8 号中国地质大学产学研基地中地大楼 A806, Guangdong 518057 (CN)。

(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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, 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, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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, MD, 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, ML, MR, NE, SN, TD, TG)。

[见续页]

(54) Title: METHOD AND DEVICE FOR OBTAINING CROSS LANGUAGE SEARCH RESOURCES AND CORRESPONDING SEARCH METHOD AND DEVICE

(54) 发明名称 : 跨语言获取搜索资源的方法和装置及对应搜索方法和装置

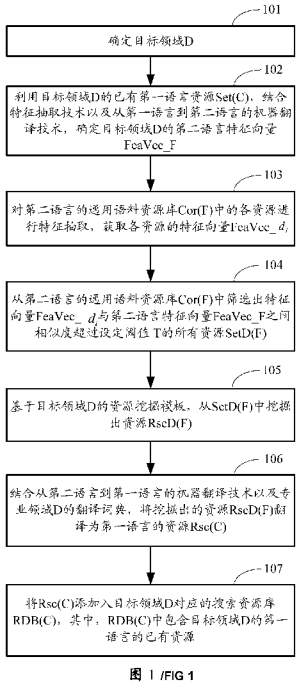


图 1 /FIG 1

101 DETERMINING A TARGET AREA D
102 UTILIZING A PREEXISTING FIRST LANGUAGE RESOURCE SET(C) IN THE TARGET AREA D, COMBINING FEATURE EXTRACTION TECHNOLOGY AND MACHINE TRANSLATION FROM THE FIRST LANGUAGE INTO A SECOND LANGUAGE TO DETERMINE A SECOND LANGUAGE FEATURE VECTOR FEAVEC_F IN THE TARGET AREA D
103 UTILIZING FEATURE EXTRACTION ON EACH RESOURCE IN A SECOND LANGUAGE UNIVERSAL CORPUS RESOURCE DATABASE COR (F) TO OBTAIN A FEATURE VECTOR FEAVEC_DI OF EACH RESOURCE
104 SELECTING, FROM THE SECOND LANGUAGE UNIVERSAL CORPUS RESOURCE DATABASE COR(F), A RESOURCE SET D(F) BETWEEN FEATURE VECTORS FEAVEC_ DI AND FEAVEC_F THAT SURPASSES A PRESET SIMILARITY THRESHOLD T
105 MINING A RESOURCE RSCD(F) FROM SETD(F), BASED ON A RESOURCE MINING TEMPLATE OF TARGET AREA D
106 COMBINING MACHINE TRANSLATION FROM THE SECOND LANGUAGE INTO THE FIRST LANGUAGE AND A SPECIALIZED FIELD D BILINGUAL DICTIONARY, TRANSLATING THE MINED RESOURCE RSCD (F) INTO THE FIRST LANGUAGE RESOURCE RSC(C)
107 ADDING RSC(C) INTO A SEARCH RESOURCE DATABASE RDB(C) CORRESPONDING TO TARGET AREA D, WHEREIN RDB(C) COMPRISES PREEXISTING RESOURCES IN THE FIRST LANGUAGE OF TARGET AREA D

(57) Abstract: Provided in the present invention are a method and a device for obtaining cross-language search resources and a corresponding search method and device. The method for obtaining search resources comprises: determining a target area D; utilizing a preexisting first language resource Set(C) in the target area D, combining feature extraction technology and machine translation from the first language in to a second language to determine a second language feature vector FeaVec_F in the target area D; using feature extraction on each resource in a second language universal corpus resource database Cor(F) to obtain a feature vector FeaVec_di of each resource; selecting from the Cor(F) a resource Set D(F) between feature vectors FeaVec_ di and FeaVec_F that surpasses a preset similarity threshold T; adding SetD(F) to a search resource database RDB(C) in the target area D. The present invention provides more, higher quality search resources for the user and improves search efficiency.

[见续页]

WO 2012/155714 A1

本国际公布:
— 包括国际检索报告(条约第 21 条(3))。

(57) 摘要:

本发明提供了一种跨语言获取搜索资源的方法和装置及对应搜索方法和装置，其中获取搜索资源的方法包括：确定目标领域 D；利用目标领域 D 的已有第一语言资源 Set(C)，结合特征抽取技术以及从第一语言到第二语言的机器翻译技术，确定目标领域 D 的第二语言特征向量 FeaVec_F；对第二语言的通用语料资源库 Cor(F)中的各资源进行特征抽取，得到各资源的特征向量 FeaVec_d_i；从 Cor(F)中筛选出特征向量 FeaVec_d_i 与 FeaVec_F 之间相似度超过预设阈值 T 的资源 SetD(F)；将 SetD(F)添加入目标领域 D 的搜索资源库 RDB(C)。通过本发明能够为用户提供更多、更优质的搜索资源，提高搜索效果。

跨语言获取搜索资源的方法和装置及对应搜索方法和装置

本申请要求了申请日为2011年05月13日，申请号为201110124378.9发明名称为“跨语言获取搜索资源的方法和装置及对应搜索方法和装置”的中国专利申请的优先权。

技术领域

本发明涉及计算机技术领域，特别涉及一种跨语言获取搜索资源的方法和装置及对应搜索方法和装置。

背景技术

框计算是2009百度技术创新大会上提出的全新技术，为用户提供基于互联网的一站式服务。用户只要在系统中输入请求，系统就能够明确识别出用户需求，并按照需求将用户的请求分配给最优的应用或资源提供商进行处理，并将处理后得到的结果返回给用户。

当框计算应用于搜索领域时，搜索系统能够识别出用户输入的搜索请求（query）的需求，例如识别出query所对应的具体领域，然后利用该query在该领域对应的资源库中进行搜索，并将搜索结果返回给用户，从而为用户提供该领域中更加专业的信息。

然而，在现有基于框计算的搜索技术中，搜索所使用的资源库通常是单语的资源库，例如，针对旅游领域，搜索所使用的资源库为中文网站提供的旅游领域的资源，但在很多情况下，单语的资源库提供的资源并不优质或者资源很少。例如，如果用户输入query“巴黎 旅游”，搜索系统识别出该query对应旅游领域，显然法文网站提供的资源可能更多且更优质，而现有搜索所使用的资源库仅为中文网站提供的旅游领域的

资源，因此搜索效果较差。

发明内容

本发明提供了一种跨语言获取搜索资源的方法和装置及对应搜索方法和装置，以便于为用户提供更多、更优质的搜索资源，提高搜索效果。

具体技术方案如下：

一种跨语言获取搜索资源的方法，该方法包括：

A、确定目标领域 D;

B、利用目标领域 D 的已有第一语言资源 Set(C)，结合特征抽取技术以及从第一语言到第二语言的机器翻译技术，确定目标领域 D 的第二语言特征向量 FeaVec_F;

C、对第二语言的通用语料资源库 Cor(F)中的各资源进行特征抽取，得到各资源的特征向量 FeaVec_d_i;

D、从所述 Cor(F)中筛选出特征向量 FeaVec_d_i与所述 FeaVec_F 之间相似度超过预设阈值 T 的资源 SetD(F);

E、将所述 SetD(F)添加入所述目标领域 D 的搜索资源库 RDB(C)。

所述步骤 B 具体包括：

B11、利用从第一语言到第二语言的机器翻译技术，将所述目标领域 D 的已有第一语言资源 Set(C)翻译成第二语言资源 Set(F);

B12、对所述 Set(F)进行特征抽取，得到所述目标领域 D 的第二语言特征向量 FeaVec_F。

或者，所述步骤 B 具体包括：

B21、对所述目标领域 D 的已有第一语言资源 Set(C)进行特征抽取，得到所述目标领域 D 的第一语言特征向量 FeaVec_C;

B22、利用从第一语言到第二语言的机器翻译技术，将所述 FeaVec_C 翻译成所述目标领域 D 的第二语言特征向量 FeaVec_F。

上述特征抽取具体包括:

S1、对被抽取资源进行基于文本的分词处理，利用预设的停用词表对分词处理后得到的词语进行过滤后得到所述被抽取资源的特征词;

S2、按照 $W(e)=tf(e,\mathrm{Set})\times\log\frac{Freq}{freq(e)}$，计算各特征词的权重值;

S3、按照各特征词的权重值进行排序，选择排在前 N1 个的特征词组成目标领域 D 的特征向量，N1 为预设的正整数;

其中，$W(e)$为特征词 e 的权重值，$f(e,\mathrm{Set})$为特征词 e 在所述被抽取资源中出现的频次；$freq(e)$为特征词 e 的绝对词频，从所述被抽取资源所对应语种的通用语料资源库中统计得到；$Freq$ 为所有特征词的绝对词频的最大值;

当所述被抽取资源为所述 Set(F)时，所述被抽取资源所对应语种为所述第二语言，所述步骤 S3 得到的特征向量为所述 FeaVec_F;

当所述被抽取资源为所述 Set(C)时，所述被抽取资源所对应语种为所述第一语言，所述步骤 S3 得到的特征向量为所述 FeaVec_C。

当所述被抽取资源为所述 Cor(F)中的各资源时，所述被抽取资源所对应语种为所述第二语言，所述步骤 S3 得到的特征向量为所述 Cor(F)中各资源的特征向量 FeaVec_d_i。

另外，所述步骤 E 具体包括:

E11、基于预设的目标领域 D 的资源挖掘模板，从所述 SetD(F)中挖掘出资源 RscD(F);

E12、结合从第二语言到第一语言的机器翻译技术，将所述 RscD(F)

翻译为第一语言的资源 Rsc(C)，将所述 Rsc(C) 添加入所述目标领域 D 的搜索资源库 RDB(C);

其中，所述资源挖掘模板包括一个以上实例构成的实例集合，所述实例包含信息类型和信息表达式中的一种或组合。

或者，所述步骤 E 具体包括:

E21、结合从第二语言到第一语言的机器翻译技术，将所述 SetD(F) 翻译为第一语言的资源 SetD(C);

E22、将所述 SetD(C) 添加入所述目标领域 D 的搜索资源库 RDB(C)。

具体地，所述翻译所使用的翻译模型由翻译短语表构成，该翻译短语表包括：第一语言短语和对应第二语言短语构成的短语对，以及，短语对的翻译分值;

所述翻译短语表中整合有所述目标领域的翻译词典。

其中，将所述目标领域的翻译词典整合到所述翻译短语表中包括:

将所述目标领域的翻译词典中的词对作为短语对逐一添加到所述翻译短语表中；其中，

如果所述翻译短语表中尚未包含被添加的词对，则将所述被添加的词对添加到所述翻译短语表中，且所述被添加的词对在所述翻译短语表中的翻译分值 $S_{PT}(ce)$ 为： $S_{PT}(ce)=S_{dict}(ce)\times rate$；如果所述翻译短语表中已经包含所述被添加的词对，则不做添加处理;

所述 $S_{dict}(ce)$ 为所述被添加的词对 ce 在所述翻译词典中的翻译分值，

$rate$ 为从所述翻译词典到所述翻译短语表的翻译分值比， $rate=\frac{\sum_{i=1}^{n} b_i}{\sum_{i=1}^{n} a_i}$，其中， a_i 为同时出现在所述翻译词典和所述翻译短语表中的词对 ce_i 在所述

翻译词典中的翻译分值，b_i为所述词对ce_i在所述翻译短语表中的翻译分值，n为同时出现在所述翻译词典和所述翻译短语表中的词对个数。

或者，将所述目标领域的翻译词典整合到所述翻译短语表中包括：

将所述目标领域的翻译词典中的词对添加到双语平行语料库中；

将所述双语平行语料库中的句对以词语为单位对齐；

对所述双语平行语料库中的句对进行短语对的抽取，基于对抽取的短语对的统计结果计算各短语对的翻译分值；

将抽取的短语对以及短语对的翻译分值构成所述翻译短语表。

较优地，所述翻译基于预设的所述目标领域D的翻译表达模板执行；

其中所述目标领域D的翻译表达式模板中包括：一个以上实例构成的实例集合；

所述实例为：由第一语言表达式和对应第二语言表达式构成的表达式对。

一种基于跨语言获取的搜索资源的搜索方法，该方法包括：

A、接收来自用户的搜索请求query；

B、对所述query进行需求识别，识别出所述query所属的领域；

C、在所识别出领域对应的搜索资源库中进行搜索；

D、将搜索结果返回给所述用户；

其中，所述搜索资源库是采用上述跨语言获取搜索资源的方法获取的。

较优地，在所述搜索结果中如果不同来源的资源所提供同一类型信息的内容发生冲突，则从内容发生冲突的资源中选择置信度最高的资源包含在所述搜索结果中，或者，按照置信度从高到低在搜索结果中对内

容发生冲突的资源进行排序。

其中，资源的置信度由以下所列置信度类型中的一种或任意组合确定：

资源的来源网站置信度、资源的来源语种置信度以及资源的翻译置信度。

资源的来源语种置信度依据资源来源的语种与所述 query 所映射的语种之间的关系确定，具体为：设置来源于所述 query 所映射语种的资源的来源语种置信度高于来源于其他语种的资源的来源语种置信度；

其中，所述 query 所映射语种的确定方法为：

提取所述 query 的特征，将所述 query 的特征与预先训练出的各语种的特征向量进行相似度计算，确定相似度超过预设的相似度阈值的语种为所述 query 所映射的语种。

所述资源的翻译置信度由在获取所述资源时使用的翻译模型中所述资源对应的翻译分值以及所述资源在语言模型中对应的分值确定。

一种跨语言获取搜索资源的装置，该装置包括：领域确定单元、目标特征确定单元、特征抽取单元、资源筛选单元和资源添加单元；

所述领域确定单元，用于确定目标领域 D；

所述目标特征确定单元，用于利用目标领域 D 的已有语言资源 Set(C)，结合特征抽取技术以及从第一语言到第二语言的机器翻译技术，确定目标领域 D 的第二语言特征向量 FeaVec_F；

所述特征抽取单元，用于将第二语言的通用语料资源库 Cor(F)中的各资源作为被抽取资源进行特征抽取，得到各资源的特征向量 FeaVec_d_i；

所述资源筛选单元，用于从所述 Cor(F)中筛选出特征向量 FeaVec_d_i

与所述 FeaVec_F 之间相似度超过预设阈值 T 的资源 SetD(F);

所述资源添加单元，用于将所述 SetD(F)添加入所述目标领域 D 的搜索资源库 RDB(C)。

其中，所述目标特征确定单元具体包括：第一翻译子单元和第一特征获取子单元;

所述第一翻译子单元，用于利用从第一语言到第二语言的机器翻译技术，将所述目标领域 D 的已有第一语言资源 Set(C)翻译成第二语言资源 Set(F)，将所述 Set(F)发送给所述特征抽取单元;

所述特征抽取单元，用于将所述 Set(F)作为被抽取资源进行特征抽取，得到所述目标领域 D 的第二语言特征向量 FeaVec_F;

所述第一特征获取子单元，用于从所述特征抽取单元获取所述目标领域 D 的第二语言特征向量 FeaVec_F。

或者，所述目标特征确定单元具体包括：第二特征获取子单元和第二翻译子单元;

所述第二特征获取子单元，用于将所述 Set(C)发送给所述特征抽取单元，从所述特征抽取单元获取所述目标领域 D 的第一语言特征向量 FeaVec_C，并提供给所述第二翻译子单元;

所述特征抽取单元，用于将所述 Set(C)作为被抽取资源进行特征抽取，得到所述目标领域 D 的第一语言特征向量 FeaVec_C;

所述第二翻译子单元，用于利用从第一语言到第二语言的机器翻译技术，将所述 FeaVec_C 翻译成所述目标领域 D 的第二语言特征向量 FeaVec_F。

上述特征抽取单元具体包括:

分词处理子单元，用于对被抽取资源进行基于文本的分词处理；

过滤处理子单元，用于利用预设的停用词表对分词处理后得到的词语进行过滤后得到所述被抽取资源的特征词；

权重计算子单元，用于按照 $W(e)=tf(e,\mathrm{Set})\times\log\frac{Freq}{freq(e)}$，计算各特征词的权重值，其中，$W(e)$为特征词 e 的权重值，$f(e,\mathrm{Set})$为特征词 e 在所述被抽取资源中出现的频次；$freq(e)$为特征词 e 的绝对词频，从所述被抽取资源所对应语种的通用语料资源库中统计得到；$Freq$ 为所有特征词的绝对词频的最大值；

向量形成子单元，用于按照各特征词的权重值进行排序，选择排在前 N1 个的特征词组成目标领域 D 的特征向量，N1 为预设的正整数；

当所述被抽取资源为所述 Set(F)时，所述被抽取资源所对应语种为所述第二语言，所述向量形成子单元形成的特征向量为所述 FeaVec_F；

当所述被抽取资源为所述 Set(C)时，所述被抽取资源所对应语种为所述第一语言，所述向量形成子单元形成的特征向量为所述 FeaVec_C。

当所述被抽取资源为所述 Cor(F)中的各资源时，所述被抽取资源所对应语种为所述第二语言，所述向量形成子单元形成的特征向量为所述 Cor(F)中各资源的特征向量 FeaVec_d_i。

具体地，所述资源添加单元包括：资源挖掘子单元、第三翻译子单元和资源添加子单元；

所述资源挖掘子单元，用于基于预设的目标领域 D 的资源挖掘模板，从所述 SetD(F)中挖掘出资源 RscD(F)；

所述第三翻译子单元，用于结合从第二语言到第一语言的机器翻译技术，将所述 RscD(F)翻译为第一语言的资源 Rsc(C)，并将所述 Rsc(C)

发送给所述资源添加子单元;

所述资源添加子单元，用于将接收到的所述 Rsc(C)添加入所述目标领域 D 的搜索资源库 RDB(C);

其中，所述资源挖掘模板包括一个以上实例构成的实例集合，所述实例包含信息类型和信息表达式中的一种或组合。

或者,所述资源添加单元包括:第四翻译子单元和资源添加子单元;

所述第四翻译子单元，用于结合从第二语言到第一语言的机器翻译技术，将所述 SetD(F)翻译为第一语言的资源 SetD(C)，并将所述 SetD(C)发送给所述资源添加子单元;

所述资源添加子单元，用于将接收到的资源添加入所述目标领域 D 的搜索资源库 RDB(C)。

具体地，所述翻译所使用的翻译模型由翻译短语表构成，该翻译短语表包括：第一语言短语和对应第二语言短语构成的短语对，以及，短语对的翻译分值;

所述翻译短语表中整合有所述目标领域的翻译词典。

此时，该装置还包括：第一翻译模型整合单元，用于将所述目标领域的翻译词典中的词对作为短语对逐一添加到所述翻译短语表中;其中，

如果所述翻译短语表中尚未包含被添加的词对，则将所述被添加的词对添加到所述翻译短语表中，且所述被添加的词对在所述翻译短语表中的翻译分值 $S_{PT}(ce)$ 为： $S_{PT}(ce)=S_{dict}(ce)\times rate$； 如果所述翻译短语表中已经包含所述被添加的词对，则不做添加处理;

所述 $S_{dict}(ce)$ 为所述被添加的词对 ce 在所述翻译词典中的翻译分值，

$rate$ 为从所述翻译词典到所述翻译短语表的翻译分值比，$rate=\frac{\sum_{i=1}^{n} b_i}{\sum_{i=1}^{n} a_i}$，其中，$a_i$ 为同时出现在所述翻译词典和所述翻译短语表中的词对 ce_i 在所述翻译词典中的翻译分值，b_i 为所述词对 ce_i 在所述翻译短语表中的翻译分值，n 为同时出现在所述翻译词典和所述翻译短语表中的词对个数。

更进一步地，该装置还包括：第二翻译模型整合单元，用于将所述目标领域的翻译词典中的词对添加到双语平行语料库中，将所述双语平行语料库中的句对以词语为单位对齐；对所述双语平行语料库中的句对进行短语对的抽取，基于对抽取的短语对的统计结果计算各短语对的翻译分值；将抽取的短语对以及短语对的翻译分值构成所述翻译短语表。

更优地，所述翻译基于预设的所述目标领域 D 的翻译表达模板执行；

其中所述目标领域 D 的翻译表达式模板中包括：一个以上实例构成的实例集合；

所述实例为：由第一语言表达式和对应第二语言表达式构成的表达式对。

一种基于跨语言获取的搜索资源的搜索装置，该搜索装置包括：用户侧交互单元、领域识别单元和搜索处理单元；

所述用户侧交互单元，用于接收来自用户的搜索请求 query，将所述搜索处理单元的搜索结果返回给所述用户；

所述领域识别单元，用于对所述 query 进行需求识别，识别出所述 query 所属的领域；

所述搜索处理单元，用于在所述领域识别单元所识别出领域对应的搜索资源库中进行搜索；

其中，所述搜索资源库是一种跨语言获取搜索资源的装置获取的。

较优地，该搜索装置还包括：冲突处理单元，用于在所述搜索结果中如果不同来源的资源所提供同一类型信息的内容发生冲突，则从内容发生冲突的资源中选择置信度最高的资源包含在所述搜索结果中，或者，按照置信度从高到低在搜索结果中对内容发生冲突的资源进行排序。

另外，该搜索装置还包括：资源置信度确定单元，用于根据以下所列置信度类型中的一种或任意组合确定资源的置信度：资源的来源网站置信度、资源的来源语种置信度以及资源的翻译置信度。

所述资源置信度确定单元依据资源来源的语种与所述 query 所映射的语种之间的关系确定资源的来源语种置信度，具体为：设置来源于所述 query 所映射语种的资源的来源语种置信度高于来源于其他语种的资源的来源语种置信度；

所述搜索装置还包括：映射语种确定单元，用于提取所述 query 的特征，将所述 query 的特征与预先训练出的各语种的特征向量进行相似度计算，确定相似度超过预设的相似度阈值的语种为所述 query 所映射的语种。

另外，资源置信度确定单元利用在获取所述资源时使用的翻译模型中所述资源对应的翻译分值以及所述资源在语言模型中对应的分值，确定所述资源的翻译置信度。

由以上技术方案可以看出，本发明通过跨语言获取搜索资源的方式，能够将属于目标领域的其他语言资源也添加到搜索资源库中，使得搜索资源库中存在属于同一领域的来自多种语种的资源，从而在执行该领域的搜索时，为用户提供更多、更优质的搜索资源，提高搜索效果。

附图说明

图 1 为本发明实施例一提供的跨语言获取搜索资源的方法流程图;

图 2 为本发明实施例二提供的搜索方法流程图;

图 3 为本发明是实例三提供的跨语言获取搜索资源的装置结构图;

图 4 中（a）和（b）分别为目标特征确定单元的两种结构图;

图 5 为本发明实施例四提供的搜索装置结构图。

具体实施方式

为了使本发明的目的、技术方案和优点更加清楚，下面结合附图和具体实施例对本发明进行详细描述。

实施例一、

图 1 为本发明实施例一提供的跨语言获取搜索资源的方法流程图，如图 1 所示，该方法可以包括以下步骤:

步骤 101: 确定目标领域 D。

框计算的特色之一就是面向各个领域针对性地提供服务，因此，在资源挖掘阶段，需要首先选定一个目标领域，例如，可以选择旅游、购物、招聘、建筑、金融等作为目标领域。在本发明中确定的该目标领域采用 D 进行标识。

步骤 102: 利用目标领域 D 的已有第一语言资源 Set(C)，结合特征抽取技术以及从第一语言到第二语言的机器翻译技术，确定目标领域 D 的第二语言特征向量 FeaVec_F。

Set(C)指的是从第一语言的通用语料资源库中已经获取的针对目标领域 D 的资源。本发明实施例中涉及的通用语料资源库指的是没有区分具体领域的所有语料资源构成的大规模资源库，Set(C)可以看作是从第

一语言的所有网页集中已经筛选出的目标领域D对应的网页构成的网页集。

本步骤的实现可以具体采用以下两种方式:

第一种方式：首先将目标领域 D 的已有第一语言资源 Set(C)翻译成第二语言资源 Set(F)，即将已有第一语言的网页集合翻译成第二语言的网页集合；然后，对第二语言资源 Set(F)进行特征抽取，得到目标领域 D 的第二语言特征向量 FeaVec_F。

第二种方式：首先对目标领域 D 的已有第一语言资源 Set(C)进行特征抽取，得到目标领域的第一语言特征向量 FeaVec_C；然后，将第一语言特征向量 FeaVec_C 翻译成第二语言特征向量 FeaVec_F。

以上两种方式均能够确定出目标领域 D 的第二语言特征向量 FeaVec_F，但第一种方式更能够充分考虑网页资源中特征词的上下文信息，具有更高的翻译质量。

具体地，在针对 Set(F)或者 Set(C)进行特征抽取时，可以采用基于词频统计的方式。

以上述第一种方式中针对 Set(F)进行特征抽取为例，首先对 Set(F)中的资源进行基于文本的分词处理，利用预设的停用词表对分词处理后得到的词语进行过滤后得到 Set(F)的特征词，其中停用词表中包含预先基于词频统计得到的表意能力很低的词语，例如一些助词、叹词、疑问词、虚词、代词等；然后按照公式（1）计算各特征词的权重值，按照权重值对各特征词进行排序，选择排在前 N1 个的特征词组成目标领域 D 的第二语言特征向量 FeaVec_F。其中，N1 为预设的正整数。

$$W(e)=tf(e,\mathrm{Set(F)})\times\log\frac{Freq_F}{freq_F(e)},\qquad(1)$$

$W(e)$为特征词e的权重值，$tf(e,\mathrm{Set(F)})$为特征词e在 Set(F)中出现的频次。$freq_F(e)$为特征词e的绝对词频，从第二语言的通用语料资源库中统计得到，该第二语言的通用语料资源库是一个大规模的不区分领域的通用语料资源库。$Freq_F$为所有特征词的绝对词频的最大值。由于绝对词频高的词太普遍，区分性差，公式（1）中体现出在 Set(F)中出现频率高的词，而弱化绝对词频高的词。

第二种方式中，针对 Set(C)进行特征抽取时，首先对 Set(C)中的资源进行基于文本的分词处理，利用预设的停用词表对分词处理后得到的词语进行过滤后得到 Set(C)的特征词；然后按照公式（2）计算各特征词的权重值，按照权重值对各特征词进行排序，选择排在前 N2 个的特征词组成目标领域 D 的第一语言特征向量 FeaVec_C。其中，N2 为预设的正整数。

$$W(e)=tf(e,\mathrm{Set(C)})\times\log\frac{Freq_C}{freq_C(e)},\qquad(2)$$

其中，$W(e)$为特征词e的权重值，$tf(e,\mathrm{Set(C)})$为特征词e在 Set(C)中出现的频次。$freq_C(e)$为特征词e的绝对词频，从第一语言的通用语料资源库中统计得到，该第一语言的通用语料资源库是一个大规模的不区分领域的通用语料资源库。$Freq_C$为所有特征词的绝对词频的最大值。同样，公式（2）中体现出在 Set(C)中出现频率高的词，而弱化绝对词频高的词。

在本发明中无论是上述第一种方式中涉及的资源的翻译，还是上述第二种方式中涉及的特征向量的翻译，通常都会涉及专业领域的知识或服务，也就会涉及专业领域的翻译词典。例如，当上述目标领域 D 为建

筑领域时，需要建筑领域的翻译词典；当上述目标领域D为金融领域时，需要金融领域的翻译词典。

首先专业领域的翻译词典可以从该专业领域的翻译目标语言（本步骤中的翻译目标语言为第二语言）的已有资源中抽取相关术语构成术语词典，然后基于数据挖掘的方法或者人工的方法对术语词典中的每个术语进行翻译，从而得到专业领域的翻译词典。

在获取到专业领域的翻译词典后，需要将专业领域的翻译词典与已有的翻译模型进行整合，采用整合后得到的翻译模型对上述第一种方式中涉及的资源进行翻译，或者对上述第二种方式中涉及的特征向量进行翻译。其中，已有的翻译模型实际上是由翻译短语表构成的。其中翻译短语表包括：第一语言短语和对应第二语言短语构成的短语对，以及短语对的翻译分值。

具体将专业领域的翻译词典与已有的翻译模型进行整合时，可以采用以下整合策略：

第一种整合策略：直接将专业领域的翻译词典添加到翻译短语表中。在将专业领域的翻译词典中的词对作为短语对添加到翻译短语表中时，如果翻译短语表中尚未包含被添加的词对时，则将被添加的词对添加到翻译短语表中，且被添加的词对在翻译短语表中的翻译分值 $S_{PT}(ce)$ 可以按照公式（3）进行计算；如果翻译短语表中已经包含被添加的词对时，则无需执行添加处理。

$$S_{PT}(ce) = S_{dict}(ce) \times rate \qquad (3)$$

其中，$S_{dict}(ce)$ 为被添加的词对 ce 在翻译词典中的翻译分值，$rate$ 为从

翻译词典到翻译短语表的翻译分值比，$rate=\frac{\sum_{i=1}^{n}b_i}{\sum_{i=1}^{n}a_i}$，其中，$a_i$为同时出现在翻译词典和翻译短语表中的词对$ce_i$在翻译词典中的翻译分值，$b_i$为词对$ce_i$在翻译短语表中的翻译分值，$n$为同时出现在翻译词典和翻译短语表中的词对个数。

第二种整合策略：将专业领域的翻译词典添加到双语平行语料库中，然后统一进行词对齐和短语抽取得到翻译短语表。

双语平行语料库中包含的是句对，翻译词典中包含的是词对，将翻译词典中的词对作为句对添加到双语平行语料库中；然后将双语平行语料库中的句对以词为单位进行对齐，即将句对中两个句子中表达同一含义的词语对齐；对双语平行语料库中的句对进行短语对的抽取，基于对抽取的短语对的统计结果计算各短语对的翻译分值，将抽取的短语对以及短语对的翻译分值构成翻译短语表。

由于翻译分值是基于短语对的统计结果的，专业领域的翻译词典中词对出现的概率较低，可能会造成计算的翻译分值较低，为了加大专业领域的翻译词典中词对的翻译分值，可以在将翻译词典中的词对添加到双语平行语料库中时，可以将翻译词典中的词对复制 N 份后，再添加到双语平行语料库中，其中 N 为预设的正整数。

步骤 103：对第二语言的通用语料资源库 Cor(F)中的各资源进行特征抽取，获取各资源的特征向量 FeaVec_d_i。

同样，对 Cor(F)各资源的特征向量的抽取方式与步骤 102 中描述的特征抽取方式相同，即针对每一个资源d_i，抽取特征词，按照公式（4）计算特征值的权重值，按照权重值对各特征词进行排序，选择排在前 N3

个的特征词构成该资源 d_i 的特征向量 FeaVec_d_i，N3 为预设的正整数。

$$W(e)=tf(e,d_i)\times\log\frac{Freq_d}{freq_d(e)}, \quad (4)$$

其中，$W(e)$ 为特征词 e 的权重值，$tf(e,d_i)$ 为特征词 e 在 d_i 中出现的频次，$freq_d(e)$ 为特征词 e 的绝对词频，从第二语言的通用语料资源库中统计得到。$Freq_d$ 为所有特征词的绝对词频的最大值。

步骤 104：从第二语言的通用语料资源库 Cor(F)中筛选出特征向量 FeaVec_d_i 与第二语言特征向量 FeaVec_F 之间相似度超过设定阈值 T 的资源 SetD(F)。

本步骤实际是从第二语言的通用语料资源库 Cor(F)中筛选出属于目标领域 D 的资源 SetD(F)作为添加入搜索资源库 RDB(C)的候选资源。在计算相似度时可以采用通用的余弦相似度的计算方式。

步骤 105：基于目标领域 D 的资源挖掘模板，从 SetD(F)中挖掘出资源 RscD(F)。

由于在筛选出的资源 SetD(F)中，可能会包含多种类型的信息，较优地，可以对其中不同类型的信息进行筛选和整理，从而形成更有序更易查询的资源。例如，在天气领域的资源中，各页面中可能会包含一些对搜索无用的广告信息、版权信息、导航信息等，可以对这些信息进行过滤；各页面中天气领域相关的信息可能会包含：地理信息、温度信息、湿度信息、风力信息、雨雪信息、时间信息等，可以对网页中的这些信息基于预设的资源挖掘模板进行筛选和整理，最终形成资源 RscD(F)。

该资源挖掘模板可以采用人工的方式或者数据挖掘的方式配置。基于目标领域 D 的资源挖掘模板包含实例集合{k1,k2,...,kn}，各实例可以包含：信息类型、信息表达式中的一种或组合。例如，天气领域可以采

用的资源挖掘模板包括：“地点信息+时间信息+温度信息”、“地点信息+时间信息+风力信息”、“地点信息+时间信息+雨雪信息”、“地点信息+时间信息+湿度信息”、“地点信息+时间信息+温度信息+风力信息+湿度信息+雨雪信息”等实例。

本步骤实际就是按照资源挖掘模板中的实例从SetD(F)中抽取资源形成RscD(F)，例如，假设天气领域的资源挖掘模板中包含的实例为“地点信息+时间信息+温度信息”以及“地点信息+时间信息+风力信息”，则SetD(F)的网页中抽取符合该实例格式的信息形成RscD(F)。

本步骤主要用于搜索资源库RDB(C)是结构化数据库时使用，如果搜索资源库RDB(C)是非结构化数据库，则也可以不执行本步骤，将步骤104筛选出的资源SetD(F)直接添加入目标领域D对应的搜索资源库RDB(C)，或者，将步骤104筛选出的资源SetD(F)经过步骤106翻译为第一语言的资源后添加入目标领域D对应的搜索资源库RDB(C)。

步骤106：结合从第二语言到第一语言的机器翻译技术，将挖掘出的资源RscD(F)翻译为第一语言的资源Rsc(C)。

本步骤中涉及的翻译与步骤102涉及的翻译原理基本相同，将第二语言的资源RscD(F)翻译为第一语言的资源Rsc(C)时，同样会涉及专业领域的知识或服务，也会涉及专业领域的翻译词典。

专业领域的翻译词典可以从该专业领域的翻译目标语言（本步骤中的翻译目标语言为第一语言）的已有资源中抽取相关术语构成术语词典，然后基于数据挖掘的方法或者人工的方法对术语词典中的每个术语进行翻译，从而得到专业领域的翻译词典。

在获取到专业领域的翻译词典后，需要将专业领域的翻译词典与已

有的翻译模型进行整合，采用整合后得到的翻译模型将第二语言的资源RscD(F)翻译为第一语言的资源 Rsc(C)。其中，已有的翻译模型实际上是由翻译短语表构成的。其中翻译短语表包括：第一语言短语和对应第二语言短语构成的短语对，以及短语对的翻译分值。

具体将专业领域的翻译词典与已有的翻译模型进行整合时，可以采用以下整合策略：

第一种整合策略：直接将专业领域的翻译词典添加到翻译短语表中。在将专业领域的翻译词典中的词对作为短语对添加到翻译短语表中时，如果翻译短语表中尚未包含被添加的词对时，则将被添加的词对添加到翻译短语表中且被添加的词对在翻译短语表中的翻译分值 $S_{PT}(ce)$ 可以按照公式（3）进行计算；如果翻译短语表中已经包含被添加的词对时，则无需进行添加处理。

$$S_{PT}(ce) = S_{dict}(ce) \times rate \qquad (3)$$

其中，$S_{dict}(ce)$ 为被添加的词对 ce 在翻译词典中的翻译分值，$rate$ 为从翻译词典到翻译短语表的翻译分值比，$rate = \frac{\sum_{i=1}^{n} b_i}{\sum_{i=1}^{n} a_i}$，其中，$a_i$ 为同时出现在翻译词典和翻译短语表中的词对 ce_i 在翻译词典中的翻译分值，b_i 为词对 ce_i 在翻译短语表中的翻译分值，n 为同时出现在翻译词典和翻译短语表中的词对个数。

第二种整合策略：将专业领域的翻译词典添加到双语平行语料库中，然后统一进行词对齐和短语抽取得到翻译短语表。

双语平行语料库中包含的是句对，翻译词典中包含的是词对，将翻译词典中的词对作为句对添加到双语平行语料库中；然后将双语平行语

料库中的句对以词为单位进行对齐，即将句对中两个句子中表达同一含义的词语对齐；对双语平行语料库中的句对进行短语对的抽取，基于对抽取的短语对的统计结果计算各短语对的翻译分值，将抽取的短语对以及短语对的翻译分值构成翻译短语表。

由于翻译分值是基于短语对的统计结果的，专业领域的翻译词典中词对出现的概率较低，可能会造成计算的翻译分值较低，为了加大专业领域的翻译词典中词对的翻译分值，可以在将翻译词典中的词对添加到双语平行语料库中时，可以将翻译词典中的词对复制 N 份后，再添加到双语平行语料库中，其中 N 为预设的正整数。

另外，除了专业领域的词汇之外，专业领域的知识和服务往往在表达形式上也有特定模式。例如，天气领域的表达形式通常是固定的，各种语言的资源所提供的天气信息在表达形式上是大同小异的。因此，可以预先挖掘出各领域的翻译表达模板，按照该翻译表达模板进行翻译。另外，需要说明的是，在步骤 102 中第一种方式所述的将 Set(C)翻译成第二语言资源 Set(F)时，也可以基于预设的目标领域 D 的翻译表达模板执行。

翻译表达模板中可以包含一个以上实例构成的实例集合 {h1,h2,...,hm}，其中每个实例都包含第一语言和第二语言的表达式对。例如，对于天气领域的一个实例可以为包含如下两个表达式的表达式对：地点信息+雨雪信息+“温度是”+最低温度+“至”+最高温度， 地点信息+“is”+雨雪信息+“with the temperature from”+最低温度+“to”+最高温度。

本步骤中可以利用目标领域 D 的翻译表达模板将挖掘出的资源

RscD(F)翻译为第一语言的资源 Rsc(C)。举个例子，假设第二语言的资源“Shanghai is rainy with temperature from -1℃ to 10℃”，按照天气领域的翻译表达模板中的实例，可以将其翻译为“上海阴雨温度是-1℃至10℃”。

其中，目标领域 D 的翻译表达模板可以采用人工的方式或者数据挖掘的方式配置。其中数据挖掘方式可以包括：预先对目标领域中翻译目标语言（本步骤中为第一语言）的已有资源进行文本分析后，抽取出其中的高频表达式，然后将该高频表达式翻译为对应的翻译源语言（本步骤中为第二语言）表达式，将第一语言表达式和第二语言表达式构成的表达式对作为翻译表达模板的一个实例。

步骤 107: 将 Rsc(C)添加入目标领域 D 对应的搜索资源库 RDB(C)，其中，RDB(C)中包含目标领域 D 的第一语言的已有资源。

通过上述流程可以将目标领域 D 中第二语言的资源以第一语言的形式添加入搜索资料库 RDB(C)中，按照同样的方式可以将其他语言，诸如第三语言、第四语言等资源以第一语言的形式添加入 RDB(C)中。例如，可以采用上述方式将建筑领域中英文的相关资源、日文的相关资源、法文的相关资源等都添加到中文的搜索资料库 RDB(C)中。

需要说明的是，上述步骤 106 是可选步骤，目的是为了更加方便使用第一语言的用户，即使用搜索资源库 RDB(C)进行搜索后返回的搜索结果均使用第一语言。也可以不执行本步骤 106，直接执行步骤 107，即直接将第二语言的资源 RscD(F)直接添加入目标领域 D 对应的搜索资源库 RDB(C)，这样在返回给用户的搜索结果中就可能包含第二语言的资源。

实施例一中所述的流程实际上是搜索系统中后台执行的操作，在此对实施例一的流程举一个实例简单说明。假设目标领域 D 为天气领域，原来搜索资料库 RDB(C)中都是来源于中文网站的资源，希望将来源于英文的、法文的以及日文的网站资源也添加到 RDB(C)中。

在此，仅以英文为例，其他语言的执行过程类似:

将天气领域已有的中文网页集 Set(C)翻译成英文后，形成英文网页集 Set(F)；然后对英文网页集 Set(F)进行特征抽取后，得到天气领域对应的英文特征向量 FeaVec_F。例如该特征向量 FeaVec_F 中可能包含有一系列特征词:

{weather,temperature,rainy,cloudy,misty,drought,storm,snowy,storm,...}。

对英文的通用语料资源库 Cor(F)中的各资源进行特征抽取，获取各资源的特征向量 FeaVec_d_i，计算各资源的特征向量 FeaVec_d_i与 FeaVec_F 之间的相关度，从 Cor(F)中筛选出相关度超过设定阈值的资源 SetD(F)。即通过相似度计算的方式从所有英文的语料资源库 Cor(F)中筛选出天气领域的资源 SetD(F)。

如果预先定义有天气领域的资源挖掘模板，可以利用资源挖掘模板从筛选出的英文资源 SetD(F)中进一步筛选和整理出不同类型的信息，例如具体的地理信息、温度信息、湿度信息、风力信息等，最终形成天气领域的英文资源 RscD(F)。

为了方便中文用户的浏览需要，可以将 RscD(F)翻译为中文，形成中文资源 Rsc(C)，将 Rsc(C)添加入天气领域的中文搜索资源库 RDB(C)中。这样，就实现了跨语言的资源获取，使 RDB(C)中包含了天气领域的来源于英文网站的资源。针对天气领域的 query，采用 RDB(C)进行的

搜索返回的搜索结果就包含有来源于英文网站的资源。

采用同样的方法，也可以使 RDB(C)中包含有来源于法文网站的资源、日文网站的资源等，从而为用户返回更多更优质的搜索结果。

至此实施例一所示流程结束。

实施例二、

图 2 为本发明实施例二提供的搜索方法流程图，该搜索方法基于通过实施例一所述方法获取的搜索资源。如图 2 所示，该方法可以包括以下步骤：

步骤 201：接收来自用户的 query。

步骤 202：对 query 进行需求识别，识别出 query 所属的领域。

步骤 203：在所识别出领域对应的搜索资源库中进行搜索。

假设识别出 query 所属的领域为实施例一中所述的目标领域 D，则在该目标领域 D 对应的搜索资源库 RDB(C)中进行搜索。由于该 RDB(C)中包含了跨语言获取的多种语言的搜索资源，因此，能够为用户提供更多且更丰富的资源。

例如，通过实施例一方式跨语言获取到了旅游领域的来自法文网站的资源，也就是说搜索资源库 RDB(C)中包含了旅游领域的来自法文网站的资源，当用户输入“巴黎 旅游”的 query 时，识别出该 query 所属领域为旅游领域，因此，在返回给用户的搜索结果中就包含有来自法文网站的资源，该资源相对于中文网站的资源来说，可能更加优质。

步骤 204：将搜索结果返回给用户。

在向用户返回搜索结果时，可能会存在这样的问题：不同来源的资源在提供同一类型信息时，同一类型信息的内容发生冲突，例如，对于

“纽约 天气”这一 query，在获取的搜索结果中，可能包含来自中国、美国和英国的天气预报网站上的资源，但三个网站的资源中提供的温度信息可能并不相同。这就需要从中甄选出置信度最高的资源，仅将内容发生冲突的资源中置信度最高的资源包含在搜索结果中，或者按照置信度从高到低在搜索结果中对内容发生冲突的资源进行排序。

在确定各资源的置信度时，可以参考以下所列置信度类型中的一种或任意组合:

1）来源网站置信度，例如中央气象台的天气预报结果置信度高于某不知名网站。

2）来源语种置信度，来源语种置信度可以依据资源来源的语种与用户 query 所映射的语种之间的关系来确定。例如对于“日本 天气”的 query 而言，该 query 所映射的语种为日文，为了反映日文网站提供的资源更准确，可以将来源于日文的资源的来源语种置信度设置的较高，将来源于其他资源的来源语种置信度设置的较低。

本发明实施例中，可以预先对各语种的已有资源进行挖掘，训练出各语种的特征向量；用户输入 query 后，提取该 query 的特征，将提取的特征与各语种的特征向量进行相似度计算，确定相似度超过预设的相似度阈值的语种为该 query 所映射的语种。

例如，对日文对应的已有资源进行挖掘，训练出特征向量包括：东京、日本、相扑、和服、小泉、樱花……。

当用户输入的源语言 query 为“日本 天气”，则提取该 query 的特征并与各语言的特征向量进行相似度计算，确定与日文的特征向量之间的相似度超过预设的相似度阈值，则确定该 query 所映射的语种为日文。

如果基于天气领域对该 query 进行搜索后，在该 query 的搜索结果中，某个资源的来源语种为日文，则设置该资源具有较高的来源语种置信度。

3）翻译置信度，这是从可读性来考虑的，翻译置信度低意味着资源的可读性不好，质量较低。其中，翻译置信度可以由资源在翻译模型中对应的翻译分值和语言模型中对应的分值确定。其中，资源在翻译模型中对应的翻译分值可以由该资源所使用的各短语对在翻译模型中的翻译分值共同确定，例如，某个资源使用了翻译模型中的多个短语对，则该资源在翻译模型中对应的翻译分值可以该资源所使用的各短语对在翻译模型中的翻译分值的平均值。语言模型中的分值是由各词语之间的搭配状况统计得到的，资源在语言模型中对应的分值也可以由该资源所使用的各词语搭配在语言模型中的分值共同确定。

综合考虑多种置信度类型时，可以将提供同一类型信息的资源按照如下公式的方式进行打分，按照打分的状况从中仅选择打分分值最高的资源包含在搜索结果中，或者，也可以将提供统一类型信息的资源都包含在搜索结果中但按照打分分值对各资源进行排序。

$$score(r_i)=\sum_{j=1}^{M}\lambda_i h_j(r_i) \qquad (6)$$

其中，$score(r_i)$为资源r_i的打分分值，即置信度，M为考虑的置信度类型数目，$h_j(r_i)$为资源r_i在第j种置信度类型的置信度。λ_i为预设的各种置信度类型的权重参数。

以上是对本发明所提供的方法进行的描述，下面通过实施例三和实施例四分别对本发明所提供的跨语言获取搜索资源的装置和搜索装置进行详细描述。

实施例三、

图 3 为本发明实施例三提供的跨语言获取搜索资源的装置结构图，如图 3 所示，该装置可以包括：领域确定单元 300、目标特征确定单元 310、特征抽取单元 320、资源筛选单元 330 和资源添加单元 340。

领域确定单元 300，用于确定目标领域 D。

目标特征确定单元 310，用于利用目标领域 D 的已有语言资源 Set(C)，结合特征抽取技术以及从第一语言到第二语言的机器翻译技术，确定目标领域 D 的第二语言特征向量 FeaVec_F。

特征抽取单元 320，用于将第二语言的通用语料资源库 Cor(F)中的各资源作为被抽取资源进行特征抽取，得到各资源的特征向量 FeaVec_^{d_i}。

资源筛选单元 330，用于从 Cor(F)中筛选出特征向量 FeaVec_d_i 与 FeaVec_F 之间相似度超过预设阈值 T 的资源 SetD(F)。

资源添加单元 340，用于将 SetD(F)添加入目标领域 D 的搜索资源库 RDB(C)。

其中，目标特征确定单元 310 可以存在两种结构：

第一种结构如图 4 中的（a）所示，可以具体包括：第一翻译子单元 401 和第一特征获取子单元 402。

第一翻译子单元 401，用于利用从第一语言到第二语言的机器翻译技术，将目标领域 D 的已有第一语言资源 Set(C)翻译成第二语言资源 Set(F)，将 Set(F)发送给特征抽取单元 320。

特征抽取单元 320，用于将 Set(F)作为被抽取资源进行特征抽取，得到目标领域 D 的第二语言特征向量 FeaVec_F。

第一特征获取子单元 402，用于从特征抽取单元 320 获取目标领域 D 的第二语言特征向量 FeaVec_F。

也就是说，第一种结构先将将目标领域 D 的已有第一语言资源 Set(C)翻译成第二语言资源 Set(F)，然后再对 Set(F) 进行特征抽取，得到目标领域 D 的第二语言特征向量 FeaVec_F。

第二种结构如图 4 中的（b）所示，目标特征确定单元 310 可以具体包括：第二特征获取子单元 411 和第二翻译子单元 412。

第二特征获取子单元 411，用于将 Set(C)发送给特征抽取单元 320，从特征抽取单元 320 获取目标领域 D 的第一语言特征向量 FeaVec_C，并提供给第二翻译子单元 412。

特征抽取单元 320，用于将 Set(C)作为被抽取资源进行特征抽取，得到目标领域 D 的第一语言特征向量 FeaVec_C。

第二翻译子单元 412，用于利用从第一语言到第二语言的机器翻译技术，将 FeaVec_C 翻译成目标领域 D 的第二语言特征向量 FeaVec_F。

也就是说，第二种结构先对 Set(C)进行特征抽取，得到目标领域 D 的第一语言特征向量 FeaVec_C，再将 FeaVec_C 翻译为第二语言特征向量 FeaVec_F。

上述的特征抽取单元 320 可以具体包括：分词处理子单元 321、过滤处理子单元 322、权重计算子单元 323 以及向量形成子单元 324。

分词处理子单元 321，用于对被抽取资源进行基于文本的分词处理。

过滤处理子单元 322，用于利用预设的停用词表对分词处理后得到的词语进行过滤后得到被抽取资源的特征词。

权重计算子单元 323，用于按照 $W(e)=tf(e,\mathrm{Set})\times\log\frac{Freq}{freq(e)}$，计算各特征词的权重值，其中，$W(e)$为特征词 e 的权重值，$f(e,\mathrm{Set})$为特征词 e 在被

抽取资源中出现的频次；$freq(e)$为特征词e的绝对词频，从被抽取资源所对应语种的通用语料资源库中统计得到；$Freq$为所有特征词的绝对词频的最大值。

向量形成子单元 324，用于按照各特征词的权重值进行排序，选择排在前 N1 个的特征词组成目标领域 D 的特征向量，N1 为预设的正整数。

其中，当被抽取资源为 Set(F)时，被抽取资源所对应语种为第二语言，向量形成子单元 324 形成的特征向量为 FeaVec_F。

当被抽取资源为 Set(C)时，被抽取资源所对应语种为第一语言，向量形成子单元 324 形成的特征向量为 FeaVec_C。

当被抽取资源为 Cor(F)中的各资源时，被抽取资源所对应语种为第二语言，向量形成子单元 324 形成的特征向量为 Cor(F)中各资源的特征向量 FeaVec_d_i。

另外，上述资源添加单元 340 可以包括：资源挖掘子单元 341 和资源添加子单元 342。

资源挖掘子单元 341，用于基于预设的目标领域 D 的资源挖掘模板，从 SetD(F)中挖掘出资源 RscD(F)，并将 RscD(F)发送给资源添加子单元 342。

资源添加子单元 342，用于将接收到的资源添加入目标领域 D 的搜索资源库 RDB(C)。

其中，资源挖掘模板可以包括一个以上实例构成的实例集合，实例包含信息类型和信息表达式中的一种或组合。

基于上述结构，资源添加单元 340 还可以包括：第三翻译子单元 343，用于获取资源挖掘子单元 341 发送给资源添加子单元 342 的 RscD(F)，

结合从第二语言到第一语言的机器翻译技术，将 RscD(F)翻译为第一语言的资源 Rsc(C)，并将 Rsc(C)发送给资源添加子单元 342。

或者，资源添加单元还可以采用另一种结构（该结构在图 3 中未示出），包括：第四翻译子单元和资源添加子单元。

第四翻译子单元，用于结合从第二语言到第一语言的机器翻译技术，将 SetD(F)翻译为第一语言的资源 SetD(C)，并将 SetD(C)发送给资源添加子单元；

资源添加子单元，用于将接收到的资源添加入目标领域 D 的搜索资源库 RDB(C)。

其中，上述第一翻译子单元 401、第二翻译子单元 412、第三翻译子单元 343 和第四翻译子单元在进行翻译所使用的翻译模型由翻译短语表构成，该翻译短语表包括：第一语言短语和对应第二语言短语构成的短语对，以及，短语对的翻译分值。其中，翻译短语表中整合有目标领域的翻译词典。

由于在专业领域中，往往会涉及专业领域的知识和服务，因此在进行翻译时，也会涉及专业领域的词汇翻译，为了达到更好的翻译效果，如上所述会将目标领域的翻译词典整合到翻译短语表中。在进行整合时可以采用实施例一中所述的两种整合方式，相应地，装置可以采用以下两种整合单元，即：

该装置还可以包括：第一翻译模型整合单元 350，用于将目标领域的翻译词典中的词对作为短语对逐一添加到翻译短语表中。

其中，如果翻译短语表中尚未包含被添加的词对，则将被添加的词对添加到翻译短语表中，且被添加词对在翻译短语表中的翻译分值

$S_{PT}(ce)$按照$S_{PT}(ce)=S_{dict}(ce)\times rate$进行计算；如果翻译短语表中已经包含被添加的词对，则无需进行添加处理。

$S_{dict}(ce)$为被添加的词对ce在翻译词典中的翻译分值，$rate$为从翻译词典到翻译短语表的翻译分值比，$rate=\frac{\sum_{i=1}^{n}b_i}{\sum_{i=1}^{n}a_i}$，其中，$a_i$为同时出现在翻译词典和翻译短语表中的词对$ce_i$在翻译词典中的翻译分值，$b_i$为词对$ce_i$在翻译短语表中的翻译分值，$n$为同时出现在翻译词典和翻译短语表中的词对个数。

或者，采用第二翻译模型整合单元（图 3 中未示出），用于将目标领域的翻译词典中的词对添加到双语平行语料库中，将双语平行语料库中的句对以词语为单位对齐；对双语平行语料库中的句对进行短语对的抽取，基于对抽取的短语对的统计结果计算各短语对的翻译分值；将抽取的短语对以及短语对的翻译分值构成翻译短语表。

由于翻译分值是基于短语对的统计结果的，专业领域的翻译词典中词对出现的概率较低，可能会造成计算的翻译分值较低，为了加大专业领域的翻译词典中词对的翻译分值，第二翻译模型整合单元可以将翻译词典中的词对复制 N 份后，再添加到双语平行语料库中，其中 N 为预设的正整数。

另外，除了专业领域的词汇之外，专业领域的知识和服务往往在表达形式上也有特定模式，为了使翻译结果能符合特定模式的习惯，上述第一翻译子单元 401、第三翻译子单元 343 和第四翻译子单元进行的翻译基于预设的目标领域 D 的翻译表达模板执行。其中目标领域 D 的翻译表达式模板中包括：一个以上实例构成的实例集合；各实例可以为：由

第一语言表达式和对应第二语言表达式构成的表达式对。

实施例四、

图 5 为本发明实施例四提供的搜索装置结构图，该装置属于搜索系统前台的装置，其对应的后台使用实施例三中所述的装置来获取搜索所使用的搜索资源。如图 5 所示，该搜索装置可以包括：用户侧交互单元 500、领域识别单元 510 和搜索处理单元 520。

用户侧交互单元 500，用于接收来自用户的 query，将搜索处理单元 520 的搜索结果返回给用户。

领域识别单元 510，用于对 query 进行需求识别，识别出 query 所属的领域。

搜索处理单元 520，用于在领域识别单元 510 所识别出领域对应的搜索资源库中进行搜索。

其中，搜索资源库是实施例三所示的装置获取的。

为了解决在搜索结果中，不同来源的资源在提供同一类型信息时，同一类型信息的内容可能发生冲突的问题，该搜索装置还可以包括：冲突处理单元 530，用于在搜索结果中如果不同来源的资源所提供同一类型信息的内容发生冲突，则从内容发生冲突的资源中选择置信度最高的资源包含在搜索结果中，或者，按照置信度从高到低在搜索结果中对内容发生冲突的资源进行排序。

该搜索装置还包括：资源置信度确定单元 540，用于根据以下所列置信度类型中的一种或任意组合确定资源的置信度：资源的来源网站置信度、资源的来源语种置信度以及资源的翻译置信度。

其中，资源置信度确定单元 540 可以按照如下公式确定资源的置信

度：

$$score(r_i)=\sum_{j=1}^{M}\lambda_i h_j(r_i)$$

其中，$score(r_i)$为资源r_i的置信度，M为置信度类型数目，$h_j(r_i)$为资源r_i在第j种置信度类型的置信度，λ_i为预设的各种置信度类型的权重参数。

在确定资源的来源语种置信度时，资源置信度确定单元 540 可以依据资源来源的语种与 query 所映射的语种之间的关系确定资源的来源语种置信度，具体为：设置来源于 query 所映射语种的资源的来源语种置信度高于来源于其他语种的资源的来源语种置信度。

此时，该搜索装置还可以包括：映射语种确定单元 550，用于提取 query 的特征，将 query 的特征与预先训练出的各语种的特征向量进行相似度计算，确定相似度超过预设的相似度阈值的语种为 query 所映射的语种。

在确定资源的翻译置信度时，资源置信度确定单元 540 可以利用在获取资源时使用的翻译模型中资源对应的翻译分值以及资源在语言模型中对应的分值，确定资源的翻译置信度。

其中，资源在翻译模型中对应的翻译分值可以由该资源所使用的各短语对在翻译模型中的翻译分值共同确定。资源在语言模型中对应的分值可以由该资源所使用的各词语搭配在语言模型中的分值共同确定。

以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，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，均应包含在本发明保护的范围之内。

权 利 要 求 书

1、一种跨语言获取搜索资源的方法，其特征在于，该方法包括:

A、确定目标领域 D;

B、利用目标领域 D 的已有第一语言资源 Set(C)，结合特征抽取技术以及从第一语言到第二语言的机器翻译技术，确定目标领域 D 的第二语言特征向量 FeaVec_F;

C、对第二语言的通用语料资源库 Cor(F)中的各资源进行特征抽取，得到各资源的特征向量 FeaVec_d_i;

D、从所述 Cor(F)中筛选出特征向量 FeaVec_d_i与所述 FeaVec_F 之间相似度超过预设阈值 T 的资源 SetD(F);

E、将所述 SetD(F)添加入所述目标领域 D 的搜索资源库 RDB(C)。

2、根据权利要求 1 所述的方法，其特征在于，所述步骤 B 具体包括：利用从第一语言到第二语言的机器翻译技术，将所述目标领域 D 的已有第一语言资源 Set(C)翻译成第二语言资源 Set(F)，对所述 Set(F)进行特征抽取，得到所述目标领域 D 的第二语言特征向量 FeaVec_F; 或者，

对所述目标领域 D 的已有第一语言资源 Set(C)进行特征抽取，得到所述目标领域 D 的第一语言特征向量 FeaVec_C，利用从第一语言到第二语言的机器翻译技术，将所述 FeaVec_C 翻译成所述目标领域 D 的第二语言特征向量 FeaVec_F。

3、根据权利要求 1 或 2 所述的方法，其特征在于，所述特征抽取具体包括:

S1、对被抽取资源进行基于文本的分词处理，利用预设的停用词表对分词处理后得到的词语进行过滤后得到所述被抽取资源的特征词;

S2、按照$W(e)=tf(e,\mathrm{Set})\times\log\frac{Freq}{freq(e)}$，计算各特征词的权重值；

S3、按照各特征词的权重值进行排序，选择排在前 N1 个的特征词组成目标领域 D 的特征向量，N1 为预设的正整数；

其中，$W(e)$为特征词 e 的权重值，$f(e,\mathrm{Set})$为特征词 e 在所述被抽取资源中出现的频次；$freq(e)$为特征词 e 的绝对词频，从所述被抽取资源所对应语种的通用语料资源库中统计得到；$Freq$ 为所有特征词的绝对词频的最大值；

当所述被抽取资源为所述 Set(F)时，所述被抽取资源所对应语种为所述第二语言，所述步骤 S3 得到的特征向量为所述 FeaVec_F；

当所述被抽取资源为所述 Set(C)时，所述被抽取资源所对应语种为所述第一语言，所述步骤 S3 得到的特征向量为所述 FeaVec_C；

当所述被抽取资源为所述 Cor(F)中的各资源时，所述被抽取资源所对应语种为所述第二语言，所述步骤 S3 得到的特征向量为所述 Cor(F)中各资源的特征向量 FeaVec_d_i。

4、根据权利要求 1 所述的方法，其特征在于，所述步骤 E 具体包括：

E11、基于预设的目标领域 D 的资源挖掘模板，从所述 SetD(F)中挖掘出资源 RscD(F)；

E12、结合从第二语言到第一语言的机器翻译技术，将所述 RscD(F)翻译为第一语言的资源 Rsc(C)，将所述 Rsc(C) 添加入所述目标领域 D 的搜索资源库 RDB(C)；

其中，所述资源挖掘模板包括一个以上实例构成的实例集合，所述实例包含信息类型和信息表达式中的一种或组合；或者，

所述步骤 E 具体包括:

E21、结合从第二语言到第一语言的机器翻译技术，将所述 SetD(F) 翻译为第一语言的资源 SetD(C);

E22、将所述 SetD(C) 添加入所述目标领域 D 的搜索资源库 RDB(C)。

5、根据权利要求 2 或 4 所述的方法，其特征在于，所述翻译所使用的翻译模型由翻译短语表构成，该翻译短语表包括：第一语言短语和对应第二语言短语构成的短语对，以及，短语对的翻译分值;

所述翻译短语表中整合有所述目标领域的翻译词典。

6、根据权利要求 5 所述的方法，其特征在于，将所述目标领域的翻译词典整合到所述翻译短语表中包括:

将所述目标领域的翻译词典中的词对作为短语对逐一添加到所述翻译短语表中；其中，

如果所述翻译短语表中尚未包含被添加的词对，则将所述被添加的词对添加到所述翻译短语表中，且所述被添加的词对在所述翻译短语表中的翻译分值 $S_{PT}(ce)$ 为： $S_{PT}(ce)=S_{dict}(ce)\times rate$；如果所述翻译短语表中已经包含所述被添加的词对，则不做添加处理;

所述 $S_{dict}(ce)$ 为所述被添加的词对 ce 在所述翻译词典中的翻译分值，

$rate$ 为从所述翻译词典到所述翻译短语表的翻译分值比， $rate=\frac{\sum_{i=1}^{n} b_i}{\sum_{i=1}^{n} a_i}$，其中， a_i 为同时出现在所述翻译词典和所述翻译短语表中的词对 ce_i 在所述翻译词典中的翻译分值， b_i 为所述词对 ce_i 在所述翻译短语表中的翻译分值， n 为同时出现在所述翻译词典和所述翻译短语表中的词对个数。

7、根据权利要求 5 所述的方法，其特征在于，将所述目标领域的翻

译词典整合到所述翻译短语表中包括:

将所述目标领域的翻译词典中的词对添加到双语平行语料库中;

将所述双语平行语料库中的句对以词语为单位对齐;

对所述双语平行语料库中的句对进行短语对的抽取，基于对抽取的短语对的统计结果计算各短语对的翻译分值;

将抽取的短语对以及短语对的翻译分值构成所述翻译短语表。

8、根据权利要求 2 或 4 所述的方法，其特征在于，所述翻译基于预设的所述目标领域 D 的翻译表达模板执行;

其中所述目标领域 D 的翻译表达式模板中包括：一个以上实例构成的实例集合;

所述实例为：由第一语言表达式和对应第二语言表达式构成的表达式对。

9、一种基于跨语言获取的搜索资源的搜索方法，其特征在于，该方法包括:

A、接收来自用户的搜索请求 query;

B、对所述 query 进行需求识别，识别出所述 query 所属的领域;

C、在所识别出领域对应的搜索资源库中进行搜索;

D、将搜索结果返回给所述用户;

其中，所述搜索资源库是采用权利要求 1 所述方法获取的。

10、根据权利要求 9 所述的方法，其特征在于，在所述搜索结果中如果不同来源的资源所提供同一类型信息的内容发生冲突，则从内容发生冲突的资源中选择置信度最高的资源包含在所述搜索结果中，或者，按照置信度从高到低在搜索结果中对内容发生冲突的资源进行排序;

其中，资源的置信度由以下所列置信度类型中的一种或任意组合确定：资源的来源网站置信度、资源的来源语种置信度以及资源的翻译置信度。

11、根据权利要求 10 所述的方法，其特征在于，资源的来源语种置信度依据资源来源的语种与所述 query 所映射的语种之间的关系确定，具体为：设置来源于所述 query 所映射语种的资源的来源语种置信度高于来源于其他语种的资源的来源语种置信度；

其中，所述 query 所映射语种的确定方法为：

提取所述 query 的特征，将所述 query 的特征与预先训练出的各语种的特征向量进行相似度计算，确定相似度超过预设的相似度阈值的语种为所述 query 所映射的语种。

12、根据权利要求 10 所述的方法，其特征在于，所述资源的翻译置信度由在获取所述资源时使用的翻译模型中所述资源对应的翻译分值以及所述资源在语言模型中对应的分值确定。

13、一种跨语言获取搜索资源的装置，其特征在于，该装置包括：领域确定单元、目标特征确定单元、特征抽取单元、资源筛选单元和资源添加单元；

所述领域确定单元，用于确定目标领域 D；

所述目标特征确定单元，用于利用目标领域 D 的已有第一语言资源 Set(C)，结合特征抽取技术以及从第一语言到第二语言的机器翻译技术，确定目标领域 D 的第二语言特征向量 FeaVec_F；

所述特征抽取单元，用于将第二语言的通用语料资源库 Cor(F)中的各资源作为被抽取资源进行特征抽取，得到各资源的特征向量 FeaVec_d_i；

所述资源筛选单元，用于从所述 Cor(F)中筛选出特征向量 FeaVec_d_i 与所述 FeaVec_F 之间相似度超过预设阈值 T 的资源 SetD(F);

所述资源添加单元，用于将所述 SetD(F)添加入所述目标领域 D 的搜索资源库 RDB(C)。

14、根据权利要求 13 所述的装置，其特征在于，所述目标特征确定单元具体包括：第一翻译子单元和第一特征获取子单元;

所述第一翻译子单元，用于利用从第一语言到第二语言的机器翻译技术，将所述目标领域 D 的已有第一语言资源 Set(C)翻译成第二语言资源 Set(F)，将所述 Set(F)发送给所述特征抽取单元;

所述特征抽取单元，用于将所述 Set(F)作为被抽取资源进行特征抽取，得到所述目标领域 D 的第二语言特征向量 FeaVec_F;

所述第一特征获取子单元，用于从所述特征抽取单元获取所述目标领域 D 的第二语言特征向量 FeaVec_F; 或者，

所述目标特征确定单元具体包括：第二特征获取子单元和第二翻译子单元;

所述第二特征获取子单元，用于将所述 Set(C)发送给所述特征抽取单元，从所述特征抽取单元获取所述目标领域 D 的第一语言特征向量 FeaVec_C，并提供给所述第二翻译子单元;

所述特征抽取单元，用于将所述 Set(C)作为被抽取资源进行特征抽取，得到所述目标领域 D 的第一语言特征向量 FeaVec_C;

所述第二翻译子单元，用于利用从第一语言到第二语言的机器翻译技术，将所述 FeaVec_C 翻译成所述目标领域 D 的第二语言特征向量 FeaVec_F。

15、根据权利要求 13 或 14 所述的装置，其特征在于，所述特征抽取单元具体包括:

分词处理子单元，用于对被抽取资源进行基于文本的分词处理;

过滤处理子单元，用于利用预设的停用词表对分词处理后得到的词语进行过滤后得到所述被抽取资源的特征词;

权重计算子单元，用于按照 $W(e)=tf(e,\mathrm{Set})\times\log\frac{Freq}{freq(e)}$，计算各特征词的权重值，其中，$W(e)$为特征词 e 的权重值，$f(e,\mathrm{Set})$为特征词 e 在所述被抽取资源中出现的频次；$freq(e)$为特征词 e 的绝对词频，从所述被抽取资源所对应语种的通用语料资源库中统计得到；$Freq$ 为所有特征词的绝对词频的最大值;

向量形成子单元，用于按照各特征词的权重值进行排序，选择排在前 N1 个的特征词组成目标领域 D 的特征向量，N1 为预设的正整数;

当所述被抽取资源为所述 Set(F)时，所述被抽取资源所对应语种为所述第二语言，所述向量形成子单元形成的特征向量为所述 FeaVec_F;

当所述被抽取资源为所述 Set(C)时，所述被抽取资源所对应语种为所述第一语言，所述向量形成子单元形成的特征向量为所述 FeaVec_C;

当所述被抽取资源为所述 Cor(F)中的各资源时，所述被抽取资源所对应语种为所述第二语言，所述向量形成子单元形成的特征向量为所述 Cor(F)中各资源的特征向量 FeaVec_d_i。

16、根据权利要求 13 所述的装置，其特征在于，所述资源添加单元包括：资源挖掘子单元、第三翻译子单元和资源添加子单元;

所述资源挖掘子单元，用于基于预设的目标领域 D 的资源挖掘模板，从所述 SetD(F)中挖掘出资源 RscD(F);

所述第三翻译子单元，用于结合从第二语言到第一语言的机器翻译技术，将所述 RscD(F)翻译为第一语言的资源 Rsc(C)，并将所述 Rsc(C)发送给所述资源添加子单元；

所述资源添加子单元，用于将接收到的所述 Rsc(C)添加入所述目标领域 D 的搜索资源库 RDB(C)；

其中，所述资源挖掘模板包括一个以上实例构成的实例集合，所述实例包含信息类型和信息表达式中的一种或组合；

或者，所述资源添加单元包括：第四翻译子单元和资源添加子单元；

所述第四翻译子单元，用于结合从第二语言到第一语言的机器翻译技术，将所述 SetD(F)翻译为第一语言的资源 SetD(C)，并将所述 SetD(C)发送给资源添加子单元；

资源添加子单元，用于将接收到的资源添加入所述目标领域 D 的搜索资源库 RDB(C)。

17、根据权利要求 14 或 16 所述的装置，其特征在于，所述翻译所使用的翻译模型由翻译短语表构成，该翻译短语表包括：第一语言短语和对应第二语言短语构成的短语对，以及，短语对的翻译分值；

所述翻译短语表中整合有所述目标领域的翻译词典。

18、根据权利要求 17 所述的装置，其特征在于，该装置还包括：第一翻译模型整合单元，用于将所述目标领域的翻译词典中的词对作为短语对逐一添加到所述翻译短语表中；其中，

如果所述翻译短语表中尚未包含被添加的词对，则将所述被添加的词对添加到所述翻译短语表中，且所述被添加的词对在所述翻译短语表中的翻译分值 $S_{PT}(ce)$ 为：$S_{PT}(ce)=S_{dict}(ce)\times rate$；如果所述翻译短语表中已

经包含所述被添加的词对，则不做添加处理；

所述 $S_{dict}(ce)$ 为所述被添加的词对 ce 在所述翻译词典中的翻译分值，

$rate$ 为从所述翻译词典到所述翻译短语表的翻译分值比，$rate=\frac{\sum_{i=1}^{n}b_i}{\sum_{i=1}^{n}a_i}$，其中，$a_i$ 为同时出现在所述翻译词典和所述翻译短语表中的词对 ce_i 在所述翻译词典中的翻译分值，b_i 为所述词对 ce_i 在所述翻译短语表中的翻译分值，n 为同时出现在所述翻译词典和所述翻译短语表中的词对个数。

19、根据权利要求 17 所述的装置，其特征在于，该装置还包括：第二翻译模型整合单元，用于将所述目标领域的翻译词典中的词对添加到双语平行语料库中，将所述双语平行语料库中的句对以词语为单位对齐；对所述双语平行语料库中的句对进行短语对的抽取，基于对抽取的短语对的统计结果计算各短语对的翻译分值；将抽取的短语对以及短语对的翻译分值构成所述翻译短语表。

20、根据权利要求 14 或 16 所述的装置，其特征在于，所述翻译基于预设的所述目标领域 D 的翻译表达模板执行；

其中所述目标领域 D 的翻译表达式模板中包括：一个以上实例构成的实例集合；

所述实例为：由第一语言表达式和对应第二语言表达式构成的表达式对。

21、一种基于跨语言获取的搜索资源的搜索装置，其特征在于，该搜索装置包括：用户侧交互单元、领域识别单元和搜索处理单元；

所述用户侧交互单元，用于接收来自用户的搜索请求 query，将所述搜索处理单元的搜索结果返回给所述用户；

所述领域识别单元，用于对所述 query 进行需求识别，识别出所述 query 所属的领域；

所述搜索处理单元，用于在所述领域识别单元所识别出领域对应的搜索资源库中进行搜索；

其中，所述搜索资源库是权利要求 13 所述装置获取的。

22、根据权利要求 21 所述的搜索装置，其特征在于，该搜索装置还包括：冲突处理单元，用于在所述搜索结果中如果不同来源的资源所提供同一类型信息的内容发生冲突，则从内容发生冲突的资源中选择置信度最高的资源包含在所述搜索结果中，或者，按照置信度从高到低在搜索结果中对内容发生冲突的资源进行排序；

资源置信度确定单元，用于根据以下所列置信度类型中的一种或任意组合确定资源的置信度：资源的来源网站置信度、资源的来源语种置信度以及资源的翻译置信度。

23、根据权利要求 22 所述的搜索装置，其特征在于，所述资源置信度确定单元依据资源来源的语种与所述 query 所映射的语种之间的关系确定资源的来源语种置信度，具体为：设置来源于所述 query 所映射语种的资源的来源语种置信度高于来源于其他语种的资源的来源语种置信度；

所述搜索装置还包括：映射语种确定单元，用于提取所述 query 的特征，将所述 query 的特征与预先训练出的各语种的特征向量进行相似度计算，确定相似度超过预设的相似度阈值的语种为所述 query 所映射的语种。

24、根据权利要求 22 所述的搜索装置，其特征在于，资源置信度确

定单元利用在获取所述资源时使用的翻译模型中所述资源对应的翻译分值以及所述资源在语言模型中对应的分值，确定所述资源的翻译置信度。

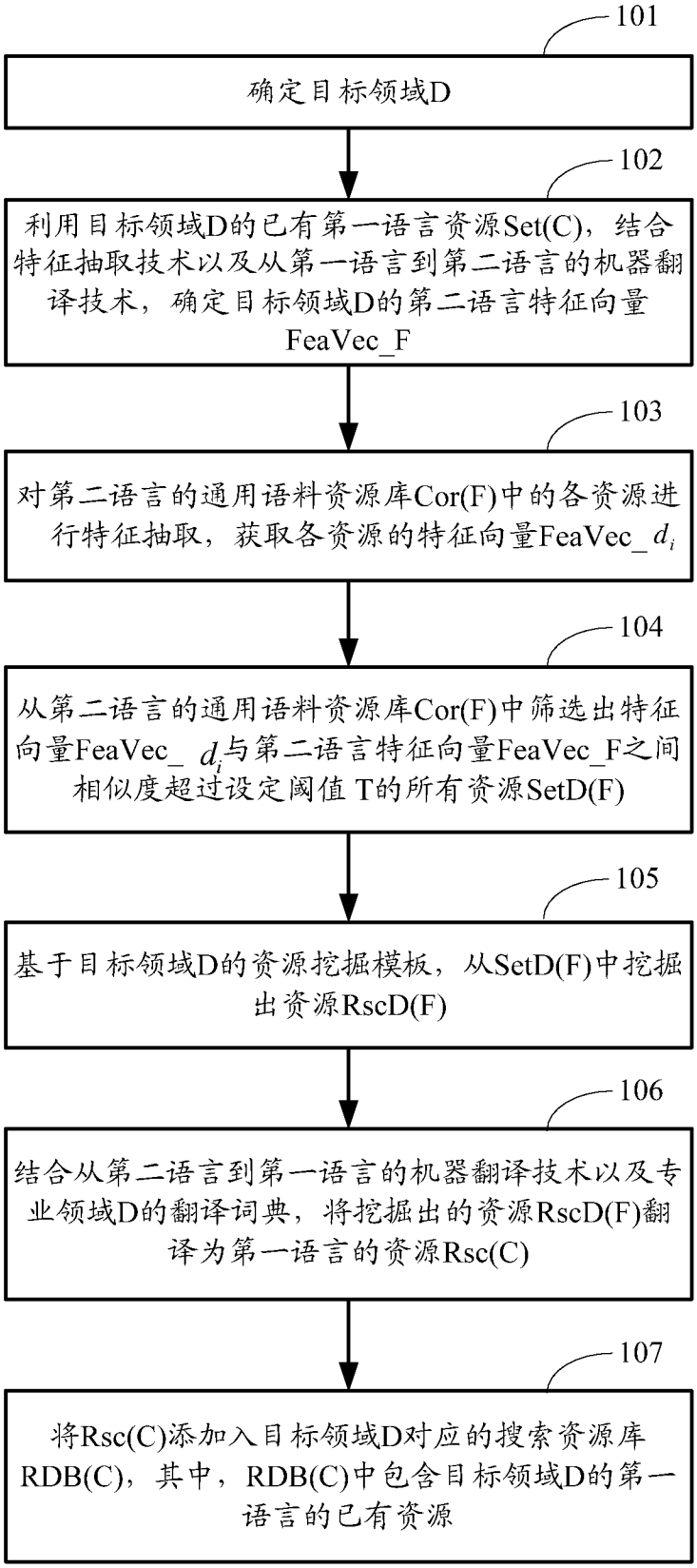

101
确定目标领域D
102
利用目标领域D的已有第一语言资源Set(C)，结合特征抽取技术以及从第一语言到第二语言的机器翻译技术，确定目标领域D的第二语言特征向量FeaVec_F
103
对第二语言的通用语料资源库Cor(F)中的各资源进行特征抽取，获取各资源的特征向量FeaVec_ d_i
104
从第二语言的通用语料资源库Cor(F)中筛选出特征向量FeaVec_ d_i与第二语言特征向量FeaVec_F之间相似度超过设定阈值 T的所有资源SetD(F)
105
基于目标领域D的资源挖掘模板，从SetD(F)中挖掘出资源RscD(F)
106
结合从第二语言到第一语言的机器翻译技术以及专业领域D的翻译词典，将挖掘出的资源RscD(F)翻译为第一语言的资源Rsc(C)
107
将Rsc(C)添加入目标领域D对应的搜索资源库RDB(C)，其中，RDB(C)中包含目标领域D的第一语言的已有资源

图 1

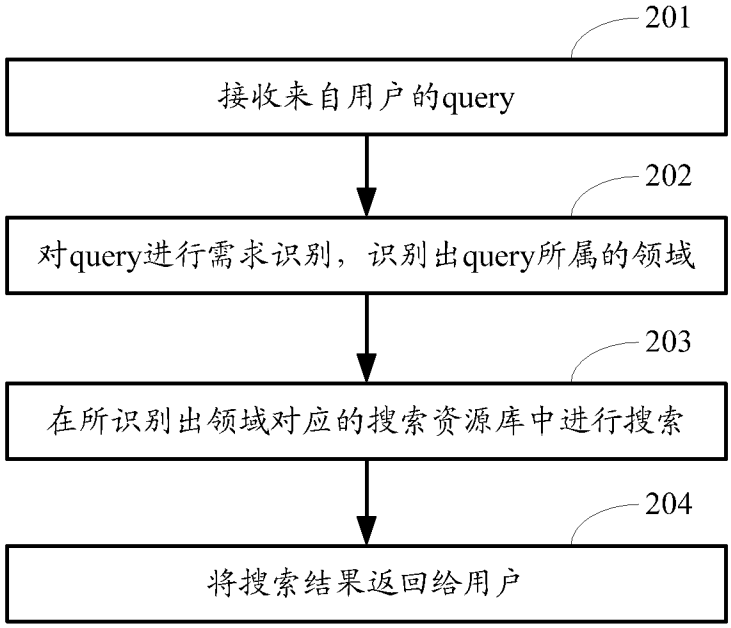

201
接收来自用户的query
202
对query进行需求识别，识别出query所属的领域
203
在所识别出领域对应的搜索资源库中进行搜索
204
将搜索结果返回给用户

图 2

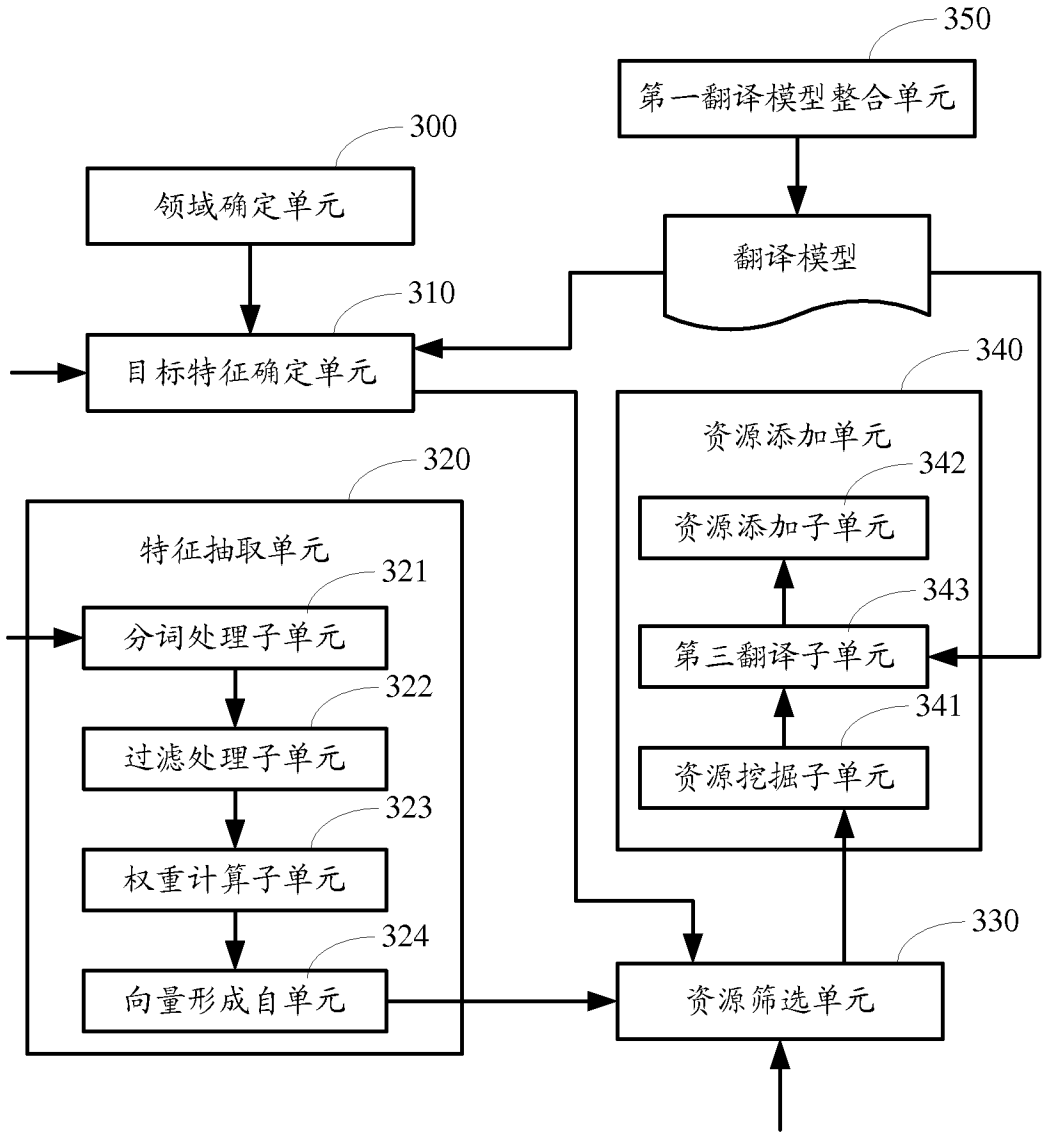
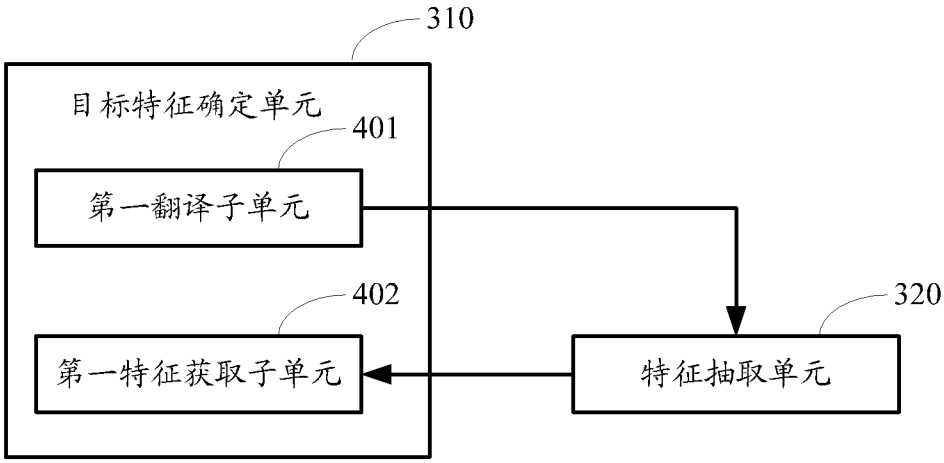
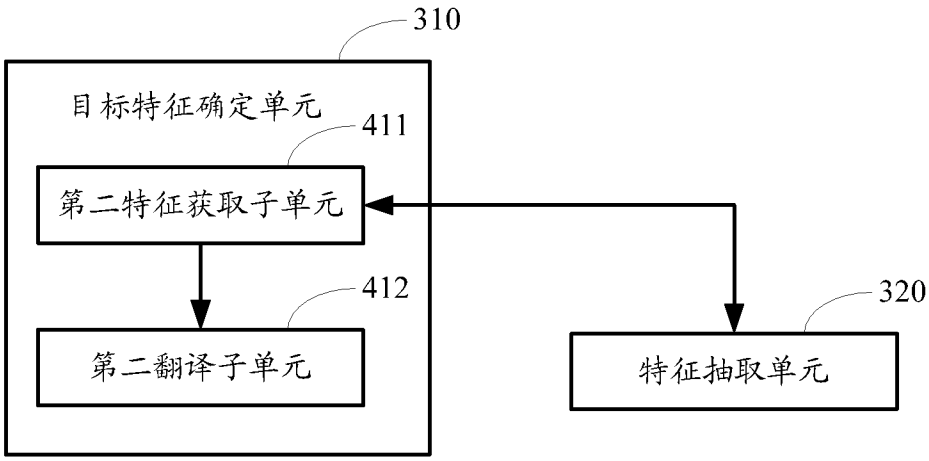

350
第一翻译模型整合单元
300
领域确定单元
翻译模型
310
目标特征确定单元
340
资源添加单元
320
特征抽取单元
342
资源添加子单元
321
分词处理子单元
343
第三翻译子单元
322
过滤处理子单元
341
资源挖掘子单元
323
权重计算子单元
330
324
向量形成自单元
资源筛选单元

图 3


310
目标特征确定单元
401
第一翻译子单元
402
第一特征获取子单元
320
特征抽取单元

（a）


310
目标特征确定单元
411
第二特征获取子单元
412
第二翻译子单元
320
特征抽取单元

（b）

图 4

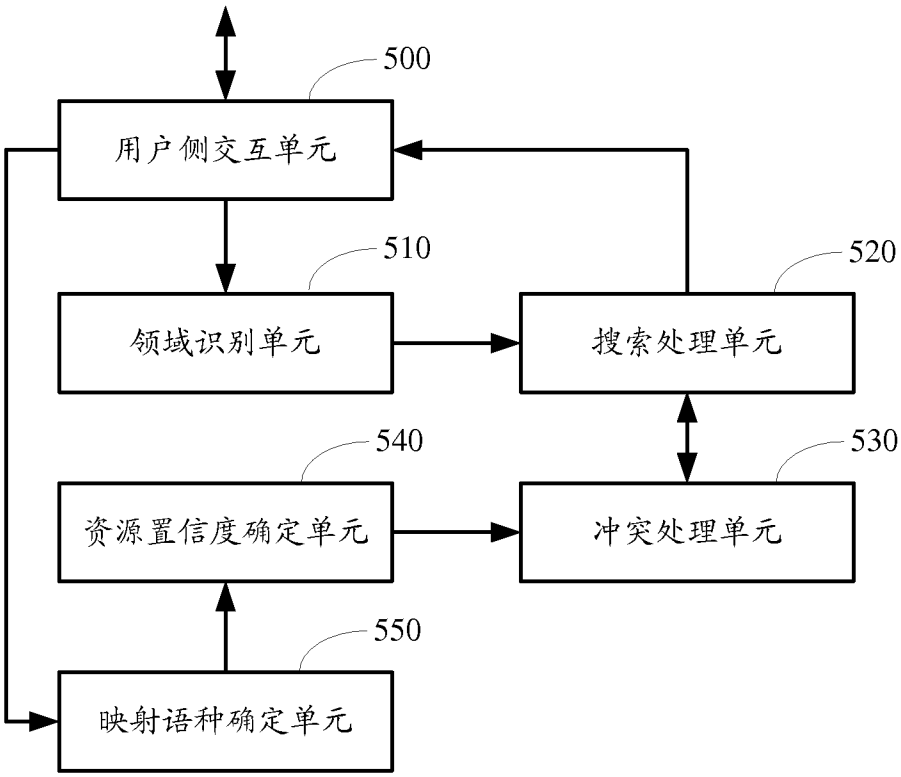

500
用户侧交互单元
510
领域识别单元
520
搜索处理单元
540
资源置信度确定单元
530
冲突处理单元
550
映射语种确定单元

图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/073656

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)

IPC: G06F /-

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)

CPRS, WPI, CNKI: BAIDU, box, library, extract, draw, cross language, CLIR, MLIR, MULTI LINGUAL, search+, quer+, retriev+, vector?, feature?, character?, database

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 101868797 A (GOOGLE INC.), 20 October 2010 (20.10.2010), description, paragraphs 17-103	1-24
A	CN 101271461 A (TOSHIBA CORPORATION), 24 September 2008 (24.09.2008), the whole document	1-24

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"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)

"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

"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

"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

"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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 June 2012 (12.06.2012)	**05 July 2012 (05.07.2012)**
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 **WANG, Nan** Telephone No.: (86-10) **62411771**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2012/073656

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 101868797 A	20.10.2010	EP 2201484 A1	30.06.2010
		US 2009193003 A1	30.07.2009
		WO 2009039524 A1	26.03.2009
		US 2009083243 A1	26.03.2009
CN 101271461 A	24.09.2008	CN 101271461 B	13.07.2011
		US 2008235202 A1	25.09.2008
		JP 2008234656 A	02.10.2008

国际检索报告

国际申请号
PCT/CN2012/073656

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)

IPC:G06F/-

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）)

CPRS,WPI,CNKI: 百度，跨语言，多语言，搜索，查询，检索，查找，框，向量，特征，库，抽取，提取，cross language，CLIR，MLIR ，MULTI LINGUAL，search+ ， quer+ ， retriev+，vector? ，feature?，character?，database

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
X	CN101868797A（谷歌公司）20.10 月 2010（20.10.2010）说明书第 17-103 段	1-24
A	CN101271461A (株式会社东芝) 24.9 月 2008（24.09.2008）全文	1－24

☐ 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	国际检索报告邮寄日期
12.6 月 2012（12.06.2012）	**05.7 月 2012 (05.07.2012)**
ISA/CN 的名称和邮寄地址： 中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号： (86-10)62019451	受权官员 王楠 电话号码：(86-10) **62411771**

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2012/073656

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN101868797A	20.10.2010	EP2201484A1	30.06.2010
		US2009193003A1	30.07.2009
		WO2009039524A1	26.03.2009
		US2009083243A1	26.03.2009
CN101271461A	24.09.2008	CN101271461B	13.07.2011
		US2008235202A1	25.09.2008
		JP2008234656A	02.10.2008