(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 101901249 A
(43)申请公布日 2010.12.01

(21)申请号 201010184725.2

(22)申请日 2010.05.12

(66)本国优先权数据
200910052055.6 2009.05.26 CN

(71)申请人 复旦大学
地址 200433 上海市邯郸路 220 号

(72)发明人 张玥杰 金城 薛向阳 岑磊 彭琳

(74)专利代理机构 上海正旦专利代理有限公司 31200
代理人 包兆宜

(51)Int.Cl.
G06F 17/30(2006.01)

权利要求书 2 页 说明书 16 页 附图 3 页

(54)发明名称

一种图像检索中基于文本的查询扩展与排序方法

(57)摘要

本发明属于多媒体信息检索领域，涉及一种在图像检索中实现基于义类词典的查询扩展与排序的方法。该发明包含：基于 WordNet 的英语词语语义相似度度量算法、基于 HowNet 的汉语词语语义相似度度量算法、基于扩展规则的查询扩展词选择与优化算法、检索结果的评分与优化算法。本发明方法使用相关的文本处理方法和语义网络词典对图像搜索引擎进行改进，通过语义扩展与用户交互及通过改进的相似度度量对检索结果进行排序。较之于传统方法而言，本发明具有准确率高、完整性强且时空代价低的优点。对于在大规模图像数据集基础上，考虑图像高层语义信息而进行高效图像检索具有非常重要的意义，在跨语言跨媒体检索领域具有广泛的应用价值。

CN 101901249 A

1. 一种图像检索中基于文本的查询扩展与排序方法，其特征在于包括如下步骤：

(1) 预处理与预分析

针对初始查询，通过预处理完成查询的分词与标点符号加标，并基于经过预处理的初始查询，通过预分析完成禁用词加标、词类分析与关键词提取；

(2) 词语语义相似度度量

针对英语词语语义相似度度量，基于网络路径长度与深度来计算语义距离，针对汉语词语语义相似度度量，基于综合考虑主类义原相似度、语义表达式相似度与主类义原框架相似度进行计算，同时融入最大匹配规则与义原深度信息；

(3) 融合扩展规则的查询扩展

基于语义网络知识，同时融合所建立的特定扩展规则，针对源于初始查询的关键词序列进行语义扩展；

(4) 基于评分的检索结果排序

以搜索引擎返回的检索结果作为处理对象，基于词语语义相似度度量评估查询关键词序列与图像描述说明之间的“相近程度”，获取评分，并通过评分算法进行优化，将最终得分作为搜索引擎返回图像的排序依据。

2. 根据权利要求 1 所述的方法，其特征在于，所述的英语词语语义相似度度量算法的原型中，建立一种基于同等词的 Lesk 扩展算法，进一步扩展词语语义定义，其中将同等词定义为某个词语所属同义词集合在 WordNet 层次结构中的兄弟结点，其中，一个同义词集合与其所对应的同等词存在一个公共父结点。

3. 根据权利要求 1 所述的方法，其特征在于，所述的汉语词语语义相似度度量算法的原型中，以整个语义表达式为基础，按层次将义原进行划分，采用最大匹配的方法，单独考虑主类义原对于概念的直接描述能力；同时，在度量过程中，加入义原深度信息的考虑，其中的概念语义相似度分为如下三个部分计算：

$Sim(C_1, C_2) = w_1*P_1+w_2*P_2+w_3*P_3$

其中，P_1 为两个概念主类义原之间的相似度；P_2 为整个语义表达式之间的相似度；P_3 是针对两个 DEF 主类义原框架之间相似度的计算；w_1、w_2 与 w_3 分别为三个部分相似度所对应的权值，应满足约束条件 $w_1+w_2+w_3 = 1$ 且 $w_2 > w_1, w_2 > w_3$。

4. 根据权利要求 1 所述的方法，其特征在于所述的融合扩展规则的查询扩展其算法步骤采用如下伪代码描述：

(1) 获得输入：原始查询关键词序列；

(2) 选择其某个关键词项；

(3) 如果为英语关键词项，查找 WordNet 的语义网络文件，获取其同义词集 Synset；

如果为汉语关键词项，查找 HowNet 的语义网络文件，获取其语义定义 DEF；

(4) 基于扩展规则，针对英语关键词项的各个 Synset，根据语义网络层次结构中的部分关系，兄弟关系，以及子女关系，寻找相应的近义词词集作为扩展词集；针对汉语关键词项的各个 DEF，作以直接匹配扩展；

(5) 基于扩展后处理策略，根据图像库标注集信息，对扩展词集进行过滤筛选，获取优化后的最终扩展词集；

(6) 重复 (2) ～ (5)，获得原始查询中每个关键词项的扩展词集进行合并，将其作为与

原始查询相对应的扩展后查询表达。

5. 根据权利要求 1 所述的方法，其特征在于，所述的基于评分的检索结果排序算法的原型中，评分算法中标注词的计算结果附加权重，用于突出图像中可能的“突出”物体；采用下述公式计算图像的排序分数：

$$Score = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} w(j,m)Sim(k_i,t_j)}{\sum_{j=1}^{m} w(j,m)}$$

其中，k_i 为关键词序列的第 i 个关键词；t_j 为图像标注词序列的第 j 个标注词；$Sim(k_i, t_j)$ 用于计算两个词项 k_i 与 t_j 之间的语义相似度；$w(j,m)$ 为相关权重，$w(j,m) = (m+1-j)^2$，用于突出标注序列中标注词项的前后关系；n 与 m 分别是查询关键词序列与图像标注词序列所包含的词项个数；当图像标注词序列中的第一个标注词权重为 m^2，则相对于总权重 $\sum_{j=1}^{m} w(j,m)$，其所占比例为：

$$\frac{m^2}{\sum_{j=1}^{m} w(j,m)} = \frac{m^2}{\sum_{j=1}^{m} j^2} = \frac{6m^2}{m(m+1)(2m+1)} = \frac{6m}{(m+1)(2m+1)}$$

该函数是一个递减函数，随着图像标注词序列的增大，排头词的权重影响成线性递减。

一种图像检索中基于文本的查询扩展与排序方法

技术领域

[0001] 本发明属于多媒体信息检索领域，涉及一种特定媒体－图像的检索方法，具体涉及一种在图像检索中实现基于义类词典的查询扩展与排序的方法。该方法可用于配合基于内容的图像检索方法，提高图像搜索质量，改善用户搜索体验。

背景技术

[0002] 近年来，随着 Internet 和社会信息化的发展，数字图像的容量正快速增长，每天都有海量的图像数据产生。如何快速而准确地查找、访问图像，并有效利用这些图像成为迫切需要解决的问题，这就是所谓的图像检索技术。最早主要采用基于文本的图像检索技术 (Text-based Image Retrieval，TBIR)；20 世纪 90 年代以来，基于内容的图像检索 (Content-based Image Retrieval，CBIR) 的研究与应用得到长足的发展，进而又发展出基于语义的图像检索、基于反馈的图像检索以及基于知识的图像检索 [1]。TBIR 作为一种早期的技术，十分依赖于图像的标注结果，这是受限之处所在，但是基于文本的检索作为一种较为成熟的技术，其快速可靠的特点至今仍然十分突出。因此，TBIR 仍然是一个值得研究的方面，如果能吸取其他方法的一些特点或是和其他几种方法交互使用，可以有不错的效果。

[0003] 从图像数据本身来看，可分为两类。一类具有相关的文本说明信息，如新闻图像，通常记者在发回图像的同时附有简短的文字描述；而另一类则是无文字说明。对于“具有文本说明信息”的图像库，一般检索系统都是从图像的文本说明中提取关键词作为索引来实现图像检索。由于多年来针对文本检索的研究已取得不少成果，与单纯基于图像底层特征的检索系统相比，这类图像检索系统往往能更好地支持基于高级语义的检索。但是，有关图像的文本说明是图像作者从自身的理解与喜好角度出发对图像所做的简短描述，与针对检索目的对图像所进行的标注信息不同，则两者所描述的内容也必然存在差异。因此，从文本说明中提取出的关键词不同于专门用于检索目的的手工标注的关键词，不仅导致查准率下降，而且可能返回用户许多不相关的查询结果，使用户无所适从 [2]。同时，在图像检索应用中，用户的真实信息需求到用户提交的查询请求之间、以及查询请求到系统理解的查询请求之间存在一定的偏差。这些偏差导致查询出来的相关图像与用户查询之间、乃至用户希望得到的信息之间存在不匹配。通过查询扩展优化用户查询，更准确、客观地表达用户的查询需求，帮助用户快速准确地获得所需要的信息，已经成为信息检索领域，特别是图像检索领域的重要研究热点 [3,4]。

[0004] 目前的查询扩展方法大致可分成三类，即基于义类词典、基于全局分析以及基于局部分析 [5,6,7]。基于义类词典的方法一般借助于语义知识词典 [8,9,10]，选择出与查询用词存在一定语义关联性的词来进行扩展，选择的依据通常为词之间的上下位关系与同义关系等 [11,12,13,14,15,16]。该方法依赖于完备的语义体系，独立于待检索对象集 [17,18,19]。基于全局分析的方法其基本思想是对全部检索对象中的词或者词组进行相关分析，将与查询用词关联程度最高的词或者词组加入初始查询以生成新的查询 [20]。该方法虽然可以最大限度地探求词间关系，但在检索对象集合改变后的更新代价巨大，而且随着检索对象集合规模的

递增在时空代价上也会存在不可行性。而基于局部分析的方法为两阶段查询，也就是首先对使用者的初始查询做第一次检索，根据检索结果选取前 N 个检索对象进行分析，找出其中重要性较高的词，与初始查询组成新的查询，然后利用新的查询进行第二次检索 [20]。该方法容易存在“查询漂移”问题，当第一次检索结果不佳时，可能会选择与查询主题不相关的词而加入至初始查询，会严重降低查询精度，甚至低于未做查询扩展的情形。

[0005] 另一方面，如何把所要的检索结果呈现给用户，帮助用户迅速地定位所需要的资源，也一直是图像检索的目标。当用户输入查询时，希望能够及时检索出最想要的结果，并且这些结果能够排在检索结果的最前面 [21]。尤其是当返回大量检索结果的时候，从用户的浏览习惯来看，基本上只关心前面若干项的结果，而很靠后的检索结果不可能也不愿意去一一遍历，甚至被用户读到的机率几乎为零 [22]。因此，检索结果的排序效果直接影响到用户能否方便地获取所需资源，同时也决定着用户对该检索系统的满意度 [23]。尤其是图像搜索引擎，其组织大量的各类图像资源，是针对特定媒体资源的信息查询工具。用户使用这类搜索引擎带有更强的目的性，更关注于能否在检索结果中尽快找到所需资源，这就对图像搜索引擎的排序处理提出更高的要求。决定排序结果的重要因素是图像搜索引擎的排序策略，而排序策略是图像搜索引擎最核心的部分之一，也是图像搜索引擎成败的关键。

[0006] 现有的通用搜索引擎排序算法从原理上可分为五种，即词频和位置加权排序算法、Direct Hit 算法、Alexa 的网站排名算法、Google 的排序算法以及相似度算法 [24,25,26]。利用词频和位置加权算法是搜索引擎早期排序的主要思想，其技术发展最成熟，至今仍是许多搜索引擎的核心排序技术。该算法的优点在于简单、易实现，比较适用于结构化文档数据。Direct Hit 是一种注重信息质量和用户行为反馈的排序算法，在一定程度上能够满足“用户保障原则”，同时也考虑信息的质量。但由于用户行为比较随意，很难保证排序结果的准确性。Alexa 专注于发布世界网站排名，主要考虑综合排名与分类排名两个方面。Google 的排序算法是其优秀搜索结果的决定性因素，采用一种精密的排序网页文件等级的方式——PageRank。查询与检索结果记录的相似程度也是搜索引擎排序的一个重要依据，目前比较常用的方法即为将查询串和文档都看作向量，其中需要考虑查询与检索结果的长度。

[0007] 通过上述的综合分析可以发现，在基于用户输入关键词进行查询的图像搜索引擎中，用户输入的关键词是引擎唯一获取的查询信息。在图像库与图像文字描述信息没有变化的条件下，与关键词序列对应的相关信息必然是唯一的检索结果。因此，从关键词序列中挖掘出尽可能多的信息来辅助查询将有助于引擎更好的理解用户的意图。查询扩展就是这样一种扩充信息的方式。如果能通过查询扩展显性信息或是隐性地缩小用户意图的不确定性，将在一定程度上能够带来更好的检索结果。另外，图像搜索引擎的查询结果往往很多，而对于用户来说，往往只会有耐心察看前几十个结果。换言之，如何将更贴近用户搜索意图的图像检索结果放到返回结果更前面的位置上相当重要。如 100 个返回结果中有 50 个正确结果不一定比只有 20 个正确结果显得更有效果，因此查准率 (Precision) 与查全率 (Recall) 都十分重要。实践显示，没有用户会把所有的搜索结果都利用起来，用户只拣有用者，这正是排序要提供给用户的便利。

[0008] 同时，以上分析也说明目前已有的查询扩展与检索结果排序算法通常源于文本检索，针对大规模文本信息处理 [27,28]。虽然基于文本的图像检索技术脱胎于现已较成熟的文

本检索技术，但其中存在某些不适用的技术，会给图像检索带来负面影响。一般的查询扩展模型或者排序模型不可能对图像检索都有效，有关现有针对文本检索的查询扩展和检索文档排序模式的研究还有待加强和深化。

[0009] 与本发明相关的参考文献有：

[0010] [1]Oilscoil Chathair，Bhaile Cliath，A.F.Sineaton，I.Quigley，Alan F.Smeaton，Ian Quigleyand Glasnevin Dublin.“Experiments on Using Semantic Distances between Words in ImageCaption Retrieval”.In Research and Development in Information Retrieval，pp.174-180，1996.

[0011] [2]Yang Linpeng，Ji Donghong，Tang Li and Niu Zhenyu.“Chinese Information Retrieval basedon Terms and Relevant Terms”.In ACM Transactions on Asian Language Information Processing；Vol.4(3)：357-374，September，2005.

[0012] [3]Xiqing Lin and Ximing Chen.New Methods for Query Expansion and Query Re-weighting forDocument Retrieval.Master Thesis，Department of Information Engineering，National Scienceand Technology University，Taiwan，2005.

[0013] [4]YiXuan Hong.Ontological Inference for User Intention Extraction，Query Expansion andConcept-based Retrieval.Master Thesis，Department of Information Engineering，NationalDong-hua University，Taiwan，2004.

[0014] [5]C.Ch.Latiri，S.Ben Yahin，J.P.ChevaVet and A.Jaouaa.“Query Expansion using FuzzyAssociation Rules between Terms”.In Proceedings of the 4th JIM International Conferenceon Knowledge Discovery and Discrete Mathematics，Mets，France，2003.

[0015] [6]Hsi-Ching Lin，Li-Hui Wang and Shyi-Ming Chen.“Query Expansion for Document Retrievalbased on Fuzzy Rules and User Relevance Feedback Techniques”.In Expert Systems withApplications，Vol.31(2)：397-405，August 2006.

[0016] [7]Hang Cui，Ji-Rong Wen，Jian-Yun Nie and Wei-Ying Ma.“Query Expansion by Mining User Logs”.In IEEE Transactions on Knowledge and Data Engineering，Vol.15(4)：829-839，July/August2003.

[0017] [8]Christiane Fellbaum(ed.).WordNet：An Electronic Lexical Database.The MIT Press，Cambridge，MA，1998.

[0018] [9]George A.Miller and Florentina Hristea.“WordNet Nouns：Classes and Instances”.InComputational Linguistics，Vol.2(1)：1-3，2006.

[0019] [10]董振东董强郝长伶，“知网的理论发现”，《中文信息学报》，Vol.21(4)：3-9，2007年。

[0020] [11]Zhiguo Gong，Chan Wa Cheang，and Leong Hou U.“Web Query Expansion by WordNet”.In LectureNotes in Computer Science Volume 4080/2006，pp.379-388，2005.

[0021] [12]Ming-Hung Hsu，Ming-Feng Tsai and Hsin-Hsi Chen.“Query Expansion with ConceptNet andWordNet：An Intrinsic Comparison”.In Proceedings of AIRS 2006，pp.1-13，2006.

[0022] [13]Alexander Budanitsky and Graeme Hirst.“Evaluating WordNet-based Measures of LexicalSemantic Relatedness”.In Computational Linguistics, Vol.32(1):13-47,2006.

[0023] [14]刘群李素建,“基于《知网》的词汇语义相似度计算”,Computational Linguistics and ChineseLanguage Processing,Vol.7(2):59-76,2002年。

[0024] [15]李峰李芳,“中文词语语义相似度计算——基于《知网》”,《中文信息学报》,Vol.21(3):99-105,2007年。

[0025] [16]江敏肖诗斌王弘蔚施水才,“一种改进的基于《知网》的词语语义相似度计算”,《中文信息学报》,Vol.22(5):84-89,2008年。

[0026] [17]Diana Inkpen and Graeme Hirst.“Building and Using a Lexical Knowledge Base of Near-SynonymDifferences”.In Computational Linguistics, Vol.32(2):223-262,2006.

[0027] [18]Ted Pedersen,Satanjeev Banerjee and Siddharth Patwardhan.“Maximizing SemanticRelatedness to Perform Word Sense Disambiguation”.In University of MinnesotaSupercomputing Institute Research Report UMSI 2005/25,March,2005.

[0028] [19]Budanitsky,Alexander and Graeme Hirst.“Semantic Distance in WordNet:An Experimental,Application-Oriented Evaluation of Five Measures.In Proceedings of the Workshop on WordNetand Other Lexical Resources,The Second Meeting of the North American Chapter of theAssociation for Computational Linguistics,Pittsburgh,PA,pp.29-34,2001.

[0029] [20]Yuen-Hsien Tseng,Da-Wei Juang and Shiu-Han Chen.“Global and Local Expansion TermExpansion for Text Retrieval”.In Proceedings of the Fourth NTCIR Workshop on Evaluationof Information Retrieval,Automatic Text Summarization and Question Answering,Tokyo,Japan,June 2-4,2004.

[0030] [21]H.Vernon Leighton,Jaideep Srivastava.Precision among World Wide Web Search Services(Search Engines):AltaVista,Excite,Hotbot,Infoseek,Lycos. September,2006.http://www.winona.msus.edu/library/webind2/webind2.htm.

[0031] [22]Claudio Carpineto,Giovanni Romano and Vittorio Giannini.“Improving Retrieval Feedbackwith Multiple Term-Ranking”.In ACM Transactions on Information Systems,Vol.20(3):259-290,July,2002.

[0032] [23]Kemafor Anyanwu,Angela Maduko and Amit Sheth.“SemRank:Ranking Complex RelationshipSearch Results on the Semantic Web”.In Proceedings of WWW 2005,pp.117-127,Chiba,Japan,May 10-14,2005.

[0033] [24]Shengli Wu and Fabio Crestani.“Methods for Ranking Information Retrieval Systems withoutRelevance Judgements”.In Proceedings of the 2003 ACM Symposium on Applied Computing,pp.811-816,Melbourne,Florida,USA,2003.

[0034] [25]Boanerges Aleman-Meza,Chris Halaschek,I.Budak Arpinar and Amit Sheth.“Context-AwareSemantic Association Ranking”.In Technical Report 03-010,

LSDIS Lab,Computer Science,Universi ty of Georgia,August,2003.

[0035] [26]Taher H.Haveliwala. “Topic-Sensitive PageRank :A Context-Sensitive Ranking Algorithmfor Web Search”.In IEEE Transactions on Knowledge and Data Engineering,Vol.15(4) :784-796,July/August,2003.

[0036] [27]Jinan Fiaidhi,Sabah Mohammed,Jihad Jaam and Ahmad Hasnah. “A Standard Framework forPersonalization via Ontology-based Query Expansion”.In Pakistan Journal of Informationand Technology,Vol.2(2) :96-103,2003.

[0037] [28]Chris Buckley and Ellen M.Voorhees. “Retrieval Evaluation with Incomplete Information”.In Proceedings of SIGIR 2004,pp.25-32,Sheffield,UK, 2004.

发明内容

[0038] 本发明的目的在于克服现有技术的不足,提出一种在基于文本的图像检索中有效进行查询扩展与检索结果排序的方法。

[0039] 本发明借鉴文本检索技术,建立一个适合于图像检索自身特点的“专用查询扩展与检索结果排序模型”。采用“从一般到特殊”的设计思想,公开一种技术框架(包括四个主要算法),即从“一般查询扩展与检索结果排序模型”出发,使用相关的文本处理方法与语义网络词典对图像搜索引擎进行改进,通过语义扩展与用户交互,其中使用语义网络词典与所建立的扩展规则;及通过改进的相似度度量对检索结果进行排序,其中重点关注于评分算法的构建与优化,最终建立适用于基于文本的图像检索的“专用查询扩展与检索结果排序模型”。

[0040] 本发明提出的图像检索中的查询扩展与检索结果排序方法,是利用相关的文本处理方法和语义网络词典对图像搜索引擎进行改进,其包含以下方面:(1) 预处理与预分析 (Pre-Processing and Pre-Analysis)——针对初始查询,通过预处理完成查询的分词与标点符号加标,并基于经过预处理的初始查询,通过预分析完成禁用词加标、词类分析与关键词提取;(2) 词语语义相似度度量 (Word Lexical Semantic Similarity Measurement)——针对英语词语语义相似度度量,基于网络路径长度与深度来计算语义距离,而针对汉语词语语义相似度度量,基于综合考虑主类义原相似度、语义表达式相似度与主类义原框架相似度进行计算,同时融入最大匹配规则与义原深度信息;(3) 融合扩展规则的查询扩展 (QueryExpansion based on Fusion of Expansion Rules)——基于语义网络知识,同时融合所建立的特定扩展规则,针对源于初始查询的关键词序列进行语义扩展;(4) 基于评分的检索结果排序 (Retrieval Results Ranking based on Scoring)——以搜索引擎返回的检索结果作为处理对象,基于词语语义相似度度量评估查询关键词序列与图像描述说明之间的“相近程度”,从而获取评分,并通过改进的评分算法进行优化,从而将最终得分作为搜索引擎返回图像的排序依据。

[0041] 与现有技术比较,本发明的上述方法在图像搜索引擎中基于所扩展的查询最终获取检索结果,具有三大优势,即准确率高、完整性强与时空代价低。其准确率高体现在查询扩展中由非常用语义词而来的扩展词非常少,这可以保证经过扩展之后所获得的扩展词与初始查询关键词具有高度共性;检索结果排序中由搜索引擎返回的相关性差或者“错误”的

图像尽可能排在检索结果列表的后面位置，这可以保证经过排序之后同一结果集中更“好”的结果排序更前，以便用户更容易看见。其完整性强体现在查询扩展中初始查询关键词序列附加扩展而来的扩展词非常完整，与图像集中的图像说明描述具有高度相关性与一致性；检索结果排序中搜索引擎的搜索行为未受任何干扰，这可以保证经过排序之后其返回结果集没有受到任何影响。其时空代价低体现在查询扩展与检索结果排序中，抛开网络传输速度与服务器处理速度等次要因素，具有更好的时空效率，而实际环境中的传输耗时将比具体计算耗时大得多，则对于用户来说感受不到这样的时间差距。

[0042] 本发明的突出贡献在于提供了 (1) 基于 WordNet 的英语词语语义相似度度量算法；(2) 基于 HowNet 的汉语词语语义相似度度量算法；(3) 基于扩展规则的查询扩展词选择与优化算法；(4) 检索结果的评分与优化算法。利用以上四个核心算法设计了一种图像检索中查询扩展与检索结果排序的技术框架。

[0043] 本发明的上述优点能满足针对大规模图像数据集，考虑图像高层语义信息而进行高效图像检索的应用需求，跨语言跨媒体检索就是其主要应用。

附图说明

[0044] 图 1 为本发明方法的流程框架图，图中标号：（一）为预处理与预分析功能模块；（二）为词语语义相似度度量功能模块；（三）为融合扩展规则的查询扩展功能模块；（四）为基于评分的检索结果排序功能模块。

[0045] 图 2 为通过具体的实例演示上述算法流程框架的具体步骤，通过给出各算法模块的中间输出以及该框架的最终检索结果，给人以直观的理解，

[0046] 其中，标号 (1) 与 (2) 分别为用户输入的原始英语查询和汉语查询；标号 (3) 与 (4) 分别为利用英语义类词典 WordNet 和汉语义类词典 HowNet，采用“基于扩展规则的查询扩展词选择与优化算法”所获取的与原始英语查询和汉语查询相对应的扩展词集；标号 (5) 与 (6) 分别为基于原始英语查询和汉语查询的扩展词集，利用图像搜索引擎所获取的相应检索结果；标号 (7) 与 (8) 分别为基于原始英语查询和汉语查询的初始检索列表，利用融合“基于 WordNet 的英语词语语义相似度度量算法”和“基于 HowNet 的汉语词语语义相似度度量算法”的“检索结果的评分与优化算法”所获取的最终检索结果。

具体实施方式

[0047] 下面结合附图详细介绍本发明在图像检索中进行查询扩展与检索结果排序的流程框架及组成该框架的四个核心算法：

[0048] 实施例 1

[0049] 1. 算法的流程框架

[0050] 附图 1 为该应用框架的流程图，标号 1-4 分别代表上述的四个主要功能模块。

[0051] 该框架分为四个主要模块：预处理与预分析 (Pre-Processing and Pre-Analysis)、词语语义相似度度量 (Concept Semantic Relativity Measurement)、融合扩展规则的查询扩展 (QueryExpansion based on Fusion of Expansion Rules) 与基于评分的检索结果排序 (Retrieval ResultsRanking based on Scoring)。四个核心算法其中的三个，基于 WordNet 的英语词语语义相似度度量算法、基于 HowNet 的汉语词语语义相

似度度量算法、以及检索结果的评分与优化算法将用于基于评分的检索结果排序模块，而基于扩展规则的查询扩展词选择与优化算法将用于融合扩展规则的查询扩模块。在该应用框架的前两个模块中，还将用到一些目前已经比较成熟的现有技术。

[0052] (1) 预处理与预分析 (Pre-Processing and Pre-Analysis)：针对初始查询，其预处理的主要任务是完成查询的分词与标点符号加标。其中，针对汉语查询采取最大概率分词策略，而针对英语查询需要附加单词首写字母大小写变换处理过程。基于经过预处理的初始查询，预分析主要完成三项任务。其一是对查询中的禁用词加以标注；其二是对每一单词进行词类分析，确定其所属正确词性，其中对于英语查询中具有变化形式的单词还需要进行形态恢复处理；而其三即为根据词类分析结果，提取作为关键词的查询词项。

[0053] (2) 词语语义相似度度量 (Word Lexical Semantic Similarity Measurement)：词语语义相似度反映是词语之间的聚合特点，可以用两个词语间的可替换程度来衡量。词语语义距离可以被看作词语语义相似度的反面，两个词语语义距离越大，则其相似度越低；反之，两个词语语义距离越小，则其相似度越大。针对英语词语语义相似度度量，基于 WordNet 义类词典的组织结构，主要考虑概念所对应的同义词集合的计算，取相似度最大（或者语义距离最小）的一对同义词集合分别代表两个同义词集合计算的最终结果，其中还需考虑针对图像搜索引擎而言时空代价的特定要求。针对汉语词语语义相似度度量，基于 HowNet 义类词典的组织结构，根据知识系统描述语言的结构特性，将语义相似度分为三部分来计算。第一部分单独考虑主类义原的相似度；第二部分考虑整个语义表达式的相似度，并根据知识描述语言描述的层次特性将义原按层次进行划分，然后每层采用最大匹配的方法进行相似度计算；第三部分考虑主类义原框架的相似度。同时，在计算义原相似度时加入义原深度信息，以区别对待含有不同信息量的义原。

[0054] (3) 融合扩展规则的查询扩展 (Query Expansion based on Fusion of ExpansionRules)：对于给定查询关键词，按照扩展规则确定另一个词作为其扩展词，即单个单词的语义扩展规则。确定该规则后，一次语义扩展就可以理解为基于语义网络知识将关键词序列的各个单词分别扩展然后将结果进行合并。考虑到用户输入的随意性，关键词序列中的单词前后顺序不作考虑，一视同仁。同时，基于图像数据集中图像说明描述信息与扩展词集规模的双重考虑，对查询扩展后关键词集进行优化处理。

[0055] (4) 基于评分的检索结果排序 (Retrieval Results Ranking based on Scoring)：基于搜索引擎返回的检索结果，根据关键字序列与图像说明描述对图像进行评分，该图像得分将作为所返回图像的排序依据。实际上，评分对象是图像说明描述，对图像本身并无任何认识，这是一个纯粹信任并依靠图像说明描述的评分方案。也就是，基于“词语语义相近”的计算支持，对查询关键词序列与图像说明描述关键词序列的“相近”程度进行评分。其中，对于图像说明描述关键词的计算结果，均赋予相应权重，用于突出图像中可能的“突出”物体。同时，通过设置适当大小的相关度计算结果缓存，建立多层缓存机制，改进评分策略，从而简化计算的复杂程度，节约大量时间，提升处理速度。

[0056] 2. 基于 WordNet 的英语词语语义相似度度量算法

[0057] 基于 WordNet 的英语词语语义相似度度量算法的创意基于以下设想：希望在在线处理图像检索结果排序过程中系统能够动态计算英语查询关键词序列与检索结果关键词序列之间的语义相似度，但是要在保证适合时空代价量级的基础上，“一般英语词语语义相

似度度量模型”的统计信息不完全丢失。即，希望利用从特定英语语义网络知识源中经过适当处理所得到的同义词集，在原词语语义相似度度量模型的基础上进行修正。然而，“一般英语词语语义相似度度量模型”存在数据稀疏与所处理词语词性受限问题，所以在线处理过程中无法融合同义词集与语义计算模型精确地度量词语语义相似度。因此，本发明提供了一种能够利用经过特殊扩展处理之后所得到同义词集在原一般英语词语语义相似度度量模型上进行修正的新算法。本发明的新算法同时满足下述的三个条件：

[0058] (1) 避免数据稀疏问题——组成词语语义定义的单词数量往往不够多，从而导致语义计算过程中发生数据稀疏问题。因此，只能利用经过特殊扩展处理之后所得到词集在原一般英语词语语义相似度度量模型上进行修正，以解决该问题。

[0059] (2) 词语词性不受限——不可以只适用于名词词语之间语义相似度的度量，应该具有跨词性词语之间的语义相似度度量能力。

[0060] (3) 时空复杂性低——处理过程中不可以使用高时空复杂性的算法，如使用概率计算的方法要读取大量的统计数据进行处理，需要考虑算法改进与优化，以满足时空代价要求。

[0061] 本发明通过下述步骤设计符合以上三个条件的新算法，

[0062] 有关英语词语语义相似度度量，经典的 Lesk 算法把词语语义定义看作为无顺序的词包，并用定义间的单词交集来衡量其相似度。Lesk 认为语义相近的词语语义定义所使用的单词也相似，但是组成定义的单词数量往往不够多，从而导致数据稀疏问题。因此，为解决这一问题，本发明提出若干扩展算法。

[0063] Lesk 扩展算法通过扩展词语语义定义，在一定程度上能够克服经典 Lesk 算法中的数据稀疏问题。EKEDAHL 和 GOLUB 通过使用 WordNet 对 Lesk 算法作以调整，通过查找某个概念的最近两个上位词，来扩充用于计算重叠个数的词语语义定义。Pedersen 等采用另一种扩展方法，考虑与某个词语在 WordNet 结构上直接相连的所有语义定义，包括上位词与下位词等，同时赋予词组更大的权重。作者宣称，在相同条件下，该算法比传统的 Lesk 算法在性能上具有显著提高。Lesk 扩展算法最常用的信息为上位词信息，即 WordNet 中的父结点，是词语语义的进一步抽象。

[0064] 现有的 Lesk 扩展算法主要考虑 WordNet 层次结构中与某个词语直接相连的语义信息，特别是上位词，来对词语语义定义进行扩展，在一定程度上能够克服数据稀疏问题。这些方法虽然能够有效利用 WordNet 结构中的直接信息，却疏忽某些非常有用的间接信息。由此，本发明建立一种基于同等词 (Coordinate Terms) 的 Lesk 扩展算法（简称 Lesk-C)，可进一步扩展词语语义定义，其中将同等词定义为某个词语所属同义词集合在 WordNet 层次结构中的兄弟结点（例如，“basketball”的同等词包括“football”、“volleyball”等）。显然，一个同义词集合与其所对应的同等词必然存在一个公共父结点。

[0065] Lesk-C 算法通过引入一个词义的所有（或者部分）同等词定义来扩展该概念语义定义，其思想基于以下假设建立，即任何概念和其同等词对于确定上下文中所起作用相一致。根据上述假设，考虑到“basketball”、“football”及“volleyball”是一组同等词，采用所建立的 Lesk-C 算法，通过使用同等词“basketball”与“football”等的定义来扩展“volleyball”的定义，无疑会增大单词相交的可能性。同等词即为 WordNet 中的兄弟结点，虽然不是原有词义的抽象甚至没有直接联系，但是在任何概念语义定义及其同等词对于确

定上下文中的概念所起作用相一致这一假设条件下，同等词和上位词同样有意义。

[0066] 对于经过基于 Lesk-C 扩展之后所获取的完整词语语义定义，由于其中每个单词都属于多个同义词集，则两个概念之间的语义相关度度量（或者语义距离计算），实际上就是两个同义词集的计算。一般来讲，采取相似度最大（或者语义距离最小）的一对同义词集分别代表两个同义词集来计算最终结果。

[0067] 本发明下文中的描述约定以及将用到的符号如下定义：

[0068] (1) 两个同义词集 S1 与 S2 在 WordNet 语义网络上的路径距离，是从 S_1 到 S_2 的路径经过的边数，用 $Len(S_1, S_2)$ 函数表示。

[0069] (2) 当只考虑 WordNet 语义网络的上下位类型的边时，语义网络退化成森林。在增加一个虚的根结点后，该森林转换为一棵树。两个同义词集 S_1 与 S_2 在上下义树里的最低公共父结点 (Lowest Super-Ordinate) 用 $Lcs(S_1, S_2)$ 函数表示，而其在树上的深度由 Depth() 函数表示。

[0070] (3) 概念语义相关度 $Sim(C_1, C_2)$ 与语义距离 $Dist(C_1, C_2)$ 之间的关系为：

[0071] $Sim(C_1, C_2)+Dist(C_1, C_2) = 1$ (1)

[0072] 在概念语义相似度度量中，把 WordNet 层次结构看成是一个图，然后利用路径信息来计算相关度。其中比较直接的想法是：两个结点的距离越近，那么两者之间的相关度越大。也就是说，如果两个结点所代表概念的公共上位词离它们越近，则这两者之间的相似度越大。这里，所使用的相似度公式如下：

[0073] $$Sim(C_1, C_2) = Sim(S_1, S_2) = \frac{2 \times Depth(Lcs(S_1, S_2))}{Depth(S_1) + Depth(S_2)} \quad (2)$$

[0074] 其中，Depth() 为概念 C 或者同义词集 S 在 WordNet 层次结构中的深度，LCS(S1, S2) 是为概念 C_1 与 C_2 或者同义词集 S_1 与 S_2 的所有公共上位词中深度最大的那个上位词。

[0075] 该公式可通过变形，转换为以下公式：

[0076]

$$Dist(C_1, C_2) = \frac{Len(C_1, Lso(C_1, C_2)) + Len(C_2, Lcs(C_1, C_2))}{Len(C_1, Lcs(C_1, C_2)) + Len(C_2, Lcs(C_1, C_2)) + 2 \times Depth(Lcs(C_1, C_2))} \quad (3)$$

[0077] 英语中存在一词多义现象，词语语义相似度应该计算概念（或者词义、语义定义）之间的相似度，两个孤立词语的语义相似度是其所有概念之间相似度的最大值。

[0078] $Sim(W_1, W_2) = \max Sim(C_{1i}, C_{2j}) i = 1 \wedge n, j = 1 \wedge m$ (4)

[0079] 其中，W1 表示词语 1 且具有 n 个概念，W_2 表示词语 2 且具有 m 个概念，C_{1i} 是 W1 的第 i 项概念，C_{2j} 是 W_2 的第 j 项概念。

[0080] 上述算法的步骤用伪代码描述如下：

[0081] (1) 获得输入：两个词语 W_1 与 W_2。

[0082] (2) 选择其两个概念 C_{1i} 与 C_{2j}。

[0083] (3) 查找 WordNet 语义网络文件，获取分别代表 C_{1i} 与 C_{2j} 的两个同义词集合 S_{1i} 与 S_{2j}。

[0084] (4) 根据公式 (1) ～ (3)，将 S_{1i} 与 S_{2j} 输入 $Dist(C_{1i}, C_{2j})$ 计算语义距离结果。

[0085] (5) 重复步骤 (2) ～ (4)，获得两个词语每一对概念之间的相似度（语义距离）值。根据公式 (4)，从中选择最大值作为最终的词语相似度值。

[0086] 其中，在计算 $Dist(C_{1i}, C_{2j})$ 时，只使用上下位关系。

[0087] 3. 基于 HowNet 的汉语词语语义相似度度量算法

[0088] 基于 HowNet 的汉语词语语义相似度度量算法的创意基于以下设想：希望在在线处理图像检索结果排序过程中系统能够动态计算汉语查询关键词序列与检索结果关键词序列之间的语义相似度，但是要在保证适合时空代价量级的基础上，能够充分考虑汉语中存在的诸多难点与复杂性。即，希望利用从特定语义网络知识源中相应的汉语词语概念多层次描述，提取丰富语义信息，建立更加符合人类主观感觉的度量机制。然而，“一般的汉语词语语义相似度度量模型”存在无法充分获取词语概念间固有关联、领域不平衡性以及数据稀疏的问题，所以在线处理过程更倾向于计算词语概念本身的相似度，而不太关注其不同语义。因此，本发明提供了一种能够利用具有“正确性，无偏见性和完备性”的词语概念语义描述在一般汉语词语语义相似度度量模型上进行修正的新算法。本发明的新算法同时满足三个条件：

[0089] (1) 避免数据稀疏问题——组成概念语义定义的单词数量往往不够多，从而导致语义计算过程中发生稀疏数据问题。因此，只能利用词语概念语义的多层次描述并附加辅助信息，从而在原一般汉语词语语义相似度度量模型上进行修正，以解决该问题。

[0090] (2) 具有高度区分力——应该能够有效利用汉语语义网络的知识结构，将不同的词语组区分在不同的相似度层次。

[0091] (3) 时空复杂性低——处理过程中不可以使用高时空复杂性的算法，如使用概率计算的方法要读取大量的统计数据进行处理，需要考虑算法改进与优化，以满足时空代价要求。

[0092] 现在描述如何设计符合以上三个条件的新算法。

[0093] 为确保词语概念语义描述的复杂度、一致性与准确性，HowNet 采用一种知识描述规范体系——知识系统描述语言 (Knowledge Database Mark-up Language，KDML)，具有以下四种重要构成形式。

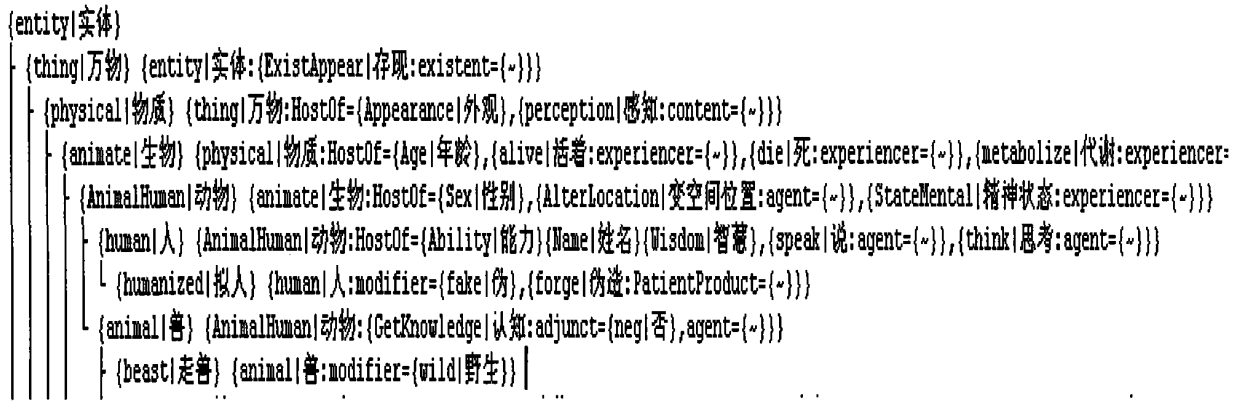
[0094] (1) 义原——KDML 中所用的词语被称为义原 (Sememes)，如“exercise| 锻炼”与“sport| 体育”，并按照 KDML 语法规则进行组织。义原不具有歧义性，是从汉字（包括单纯词）中所提取出来的“最基本且不易于再分割的意义最小单位”，也就是描述的最小单位。

[0095] (2) 主类义原——语义表达式中的第一个义原同时也被称为主类义原，前述实例中“exercise| 锻炼”即为主类义原。主类义原必须指出概念最基本的意义，可认为其对概念具有最强的描述能力。

[0096] (3) 语义表达式——“DEF ＝ {...}”是整个记录的核心，是对于概念的定义和描述，称之为语义表达式。为确保概念描述的复杂度、一致性和准确性，利用 KDML 进行规范。

[0097] (4) 主类义原框架——简单地说，就是对于大部分义原也像词语一样进行语义表达式定义，如下图所示。其中，对于义原“thing| 万物”，其主类义原框架为“{entity| 实体：{ExistAppear| 存现 :existent ＝ { ～ }}}”，描述语法严格遵循 KDML 描述语言。

[0098]



[0099] 在基于 KDML 所建立的词语概念语义描述中，处于不同括号层次中的义原对于词语语义定义的描述能力不同，越是处于外层括号中的义原对概念的描述能力越强；反之，处于内层括号中的义原是对上一层义原的具体解释，是对概念的间接描述，描述能力相对较弱。因此，在度量词语语义相似度时，有必要将其区别对待。

[0100] 作为词语相似度度量的重要基础，义原相似度的计算依据义原的层次体系（即上下位关系）进行。基于树状层次结构，考虑结点之间的路径长度，同时引入结点的层次深度，而建立义原相似度的计算公式，如下所示。

[0101] $$Sim(S_1,S_2)=\frac{\alpha\times\min(Depth(S_1),Depth(S_2))}{\alpha\times\min(Depth(S_1),Depth(S_2))+Dist(S_1,S_2)} \quad (1)$$

[0102] 其中，S_1 与 S_2 分别表示两个义原；$Dist(S_1,S_2)$ 表示义原 S_1 与 S_2 之间的路径长度；α 为调节参数；$Depth(S_1)$ 与 $Depth(S_2)$ 分别表示义原 S_1 与 S_2 的层次深度；$\min(Depth(S_1),Depth(S_2))$ 表示义原 S_1 与 S_2 层次深度中的较小者。义原所携带的语义信息具有大小之分，越是处于底层的结点语义信息越丰富，越是处于高层的结点语义越抽象，所以应该区别对待不同层次上的义原。

[0103] 汉语中存在一词多义现象，词语语义相似度应该计算词语概念之间的相似度，两个孤立词语（不处在一定的上下文背景中）的语义相似度是其所有概念之间相似度的最大值。

[0104] $Sim(W_1,W_2)=\max Sim(C_{1i},C_{2j})\ i=1\wedge n,\ j=1\wedge m$ (2)

[0105] 其中，词 W_1 具有 n 个概念，词 W2 具有 m 个概念，C_{1i} 是 W_1 的第 i 项概念，C_{2j} 是 W_2 的第 j 项概念。根据 KDML 的结构特性，将概念语义相似度分为三个部分进行计算：

[0106] $Sim(C_1,C_2)=w_1{}^*P_1+w_2{}^*P_2+w_3{}^*P_3$ (3)

[0107] 其中，P_1 为两个概念主类义原之间的相似度；P_2 为整个语义表达式之间的相似度；P_3 是针对两个 DEF 主类义原框架之间相似度的计算；w_1、w_2 与 w_3 分别为三个部分相似度所对应的权值，应满足约束条件 $w_1+w_2+w_3=1$ 且 $w_2>w_1$，$w_2>w_3$。

[0108] 对于 P_1，按公式 (1) 进行计算，前述已说明主类义原对于概念具有最直接的语义描述能力，因此将其单列为一部分进行考虑很有意义。

[0109] 对于 P_2，由于语义表达式是一个完整的个体，并拥有自己的语法规则，因此将其作为一个整体并参考 KDML 规则来计算其语义相似度很有必要。该部分是整个语义相似度度量中最复杂且权值比重最大的一部分。因为需要考虑整个语义表达式。其计算过程可分为两个阶段，根据 KDML 描述的层次特性将义原按层次进行划分，然后每层采用最大匹配的方

法进行语义相似度计算。首先，计算每组义原的语义相似度，从中选择值最大的一组。如果存在多组义原语义相似度相同，则任选一组即可。其次，在剩下的义原组中仍选择语义相似度最大者，依此类推。当两个概念同层的义原个数不等时，会出现义原和空元素配对的情况，此时可统一取较小值 r（所设定的参数）。最后，将所选出的义原组语义相似度相加取平均值，即可得到 P_2 部分的值。

[0110] 对于 P_3，其计算方法与 P_2 相同。针对主类义原框架的语义相似度度量实际上是另一种计算主类义原语义相似度的方法，再一次强调主类义原对于概念的直接描述能力。

[0111] 最终，基于上述三部分相似度的计算，根据公式 (3) 即可计算出每对概念之间的语义相似度，然后按公式 (2) 取最大值作为词语间的语义相似度。

[0112] 需要注意的一些特殊情形是，当仅用一个义原就能完全解释一个词语时，说明该义原含有的信息量比较大，是处于义原树中较底层的一个。此时，如果加入义原深度信息，则可提高该单一义原的描述能力，使词语语义相似度更接近于期望值。另外，对于使用引号括起来的特殊意义义原，也可称之为具体词，包含较丰富且具体的语义信息，对其所描述概念的性质具有直接的决定作用与影响力。因此，应该将其区别于普通义原，给具体词语之间的语义相似度赋予一个调节参数。

[0113] 在上述的语义相似度度量模型中，以整个语义表达式为基础，按层次将义原进行划分，并采用最大匹配的方法，同时单独考虑主类义原对于概念的直接描述能力。这种度量语义相似度的机制可更为有效地利用 HowNet 的知识结构，使得结果更为具有区分力。同时，由于在度量过程中，适当加入义原深度信息的考虑，使结果更加精确，尤其是在语义表达式中义原个数不多的情况下效果更加明显。

[0114] 该算法的步骤用伪代码描述如下：

[0115] (1) 获得输入：两个词语 W_1 与 W_2。

[0116] (2) 选择其两个概念 C_{1i} 与 C_{2j}。

[0117] (3) 查找 HowNet 的语义网络文件，获取概念 C_{1i} 与 C_{2j} 的主类义原、语义表达式、语义表达式框架等相关信息。

[0118] (4) 基于义原相似度的计算公式 (1)，获取两个概念主类义原之间的相似度信息 P_1。

[0119] (5) 基于两阶段的求解过程，分别计算两个概念语义表达式和主类义原框架之间的相似度 P_2 与 P_3。

[0120] (6) 综合三部分的相似度信息，根据公式 (3)，获取两个概念之间的相似度取值。

[0121] (7) 重复步骤 (2) ～ (6)，获得两个词语每一对概念之间的相似度值。根据公式 (2)，从中选择最大值作为最终的词语相似度值。

[0122] 4. 基于扩展规则的查询扩展词选择与优化算法

[0123] 针对基于义类词典语义网络所进行的查询语义扩展，有两种方式可借鉴。其一是将基于原始查询的搜索结果自动加入原始查询关键词序列中，该方式一般需要人工参与和一定规模的机器学习及积累，否则将引入大量无关词汇，使得扩展结果十分糟糕。而另一种方式是将选择权交给用户，仅提供扩展后的结果，至于是否适用或者使用哪些扩展结果则由用户决定。虽然使用后一种方式需要用户主动参与，在一定程度上增加用户使用搜索引擎的复杂度，但由此得到的扩展词实际上是一种用户输入，具有较高使用价值。

[0124] 文本检索领域中的自动查询扩展是一项较为成熟的技术，如 [3,4,7]，该类算法考虑更多的是合并检索文档中的相关信息，然而许多检索文档与查询并无关系。有关结合用户交互的半自动查询扩展的研究也比较成熟，如 [5,6,20]，该类算法通常将从检索文档中能够提取出的所有相关词汇信息全部提供给用户，造成用户面临范围宽泛的诸多选择，而容易造成选择不适当或者引入不必要的噪声信息。同时，上述两种查询扩展方式均针对文本检索并结合文本信息的特点而建立，对于基于文本的图像检索来说并非完全适用。鉴此，一种结合两种经典查询扩展技术且适用于图像检索的新算法应运而生，更容易实现与使用，是一种轻量且具有更直接效果的方法。

[0125] 确定查询扩展模式之后，在实现查询扩展功能中，首先必须考虑适合图像检索特点的查询扩展规则构建。即，对于一个给定查询关键词项，按照何种具体规则来确定另一词项是其合适的扩展词，也就是单个词项的语义扩展规则。基于扩展规则的确定，查询语义扩展就可以理解为将查询关键词序列的各个词项分别扩展然后将其合并。考虑到用户输入的随意性，关键词序列中各词项前后顺序没有差异，一视同仁。

[0126] 为最大限度地满足用户希望通过选择扩展词项而使搜索意图更为明显的目的，本发明考虑以下两种情况建立查询扩展规则。

[0127] (1) 用户对作为搜索对象的图像找不到很好的词汇进行抽象描述，另一方面图像的标注者往往使用直接而且常用的词语，使得用户的词语输入比较棘手。从而，针对图像本身的标注信息以及检索的要求，应该扩展出一些与查询关键词具有共性的词项。例如，用户希望搜索有关大型猫科动物的图像，如果输入“big_cat”，结果会很糟糕。因为大多数图像的标注信息为“tiger”、“lion”等具体词项，则返回结果就会很少。针对该情形，其最佳解决途径是，将“tiger”、“lion”等大型猫科动物的名字都输入于搜索框内，但对用户来说这种方式显然很繁琐。因此，如果能够仅输入“tiger”然后扩展出一些其它大型猫科动物的名字，则用户只需要将扩展词项选进搜索框即可，而无需多加思考还有其它哪些名字需要通过键盘输入。

[0128] (2) 作为用户输入的查询关键词项具有多种涵义的时候（这是经常出现的情况），通过选取扩展词项加入关键词序列，可谓给搜索引擎提供一定的消歧依据。

[0129] 例如，对于关键词字“bank”，如果能和“water”或者“coast”在一起，图像搜索引擎就拥有依据避免把关于“银行”的图像返回或者评分过高。

[0130] 基于上述扩展规则，在查询语义扩展中，通过搜索义类词典（包括针对英语查询的 WordNet、针对汉语查询的 HowNet）的语义网络，将与原始英语查询关键词项具有部分关系 (Part)、兄弟关系 (Sibling) 以及子女关系 (Child) 的相关词项作为扩展词项返回，而直接使用 DEF 匹配对原始汉语查询进行扩展。其中，针对英语查询扩展的子女关系仅包含直接子女，即语义层次关系中的直接子结点。

[0131] 除上述扩展规则之外，本发明还考虑扩展词项最终是要加至关键词序列中，而关键词序列的词项要用于搜索过程中与图像库的图像标注信息进行匹配处理。因此，图像标注信息中未出现过的词项由于对搜索结果无用，则在扩展模块中扩展出来毫无意义。义类词典中的词项数目通常数万，甚至十万以上，均有可能通过上述扩展规则而被选中作为扩展词项。但是，图像标注信息一般仅会出现常用词，而常用词集合就小很多。因此，在基于扩展规则的查询语义扩展中，其最后一步是利用标注词过滤扩展词集，未在标注词集中出

现的扩展词项将被抛弃。

[0132] 另外，在基于 WordNet 的英语查询语义扩展中需要解决的一个问题是，由于关键词序列中的各个词项相互独立，最终的扩展结果是各个词项扩展词集的并集。相似地，每个词项的扩展词集也可通过包含该词项各个同义词集 Synset 的扩展词集的并集得到。如果一个词项的某个 Synset 在语义网络中所处位置比较“密集”，且语义关系比较复杂，则由该 Synset 出发得到的扩展词集规模大大超出其它几个 Synset。而一个词项的所有 Synset 都具有同等地位，则有可能通过前述带来较大规模扩展集的 Synset，引入许多成为噪音的扩展词项信息。例如，在对关键词项“tiger”进行扩展时，大量的扩展词项竟然是与“人文”相关。这个意外的结果来自“tiger”的一个 Synset，其语义为“a fierce or audacious person”，而该“拟人”的语义并非“tiger”的常用语义，在未进行消歧处理的情况下，无法通过具体规则找出此类 Synset。因此，在进行语义扩展时，对每个 Synset 的扩展词集作以规模上的限制（如限制每个 Synset 至多扩展出 15 个扩展词项），从而避免因为某个比较冷僻的 Synset 扩展出大量不合格的扩展词项。

[0133] 该算法的步骤用伪代码描述如下：

[0134] (1) 获得输入：原始查询关键词序列。

[0135] (2) 选择其某个关键词项。

[0136] (3) 如果为英语关键词项，查找 WordNet 的语义网络文件，获取其同义词集 Synset。如果为汉语关键词项，查找 HowNet 的语义网络文件，获取其语义定义 DEF。

[0137] (4) 基于扩展规则，针对英语关键词项的各个 Synset，根据语义网络层次结构中的部分关系 (Part)、兄弟关系 (Sibling) 以及子女关系 (Child)，寻找相应的近义词词集作为扩展词集；针对汉语关键词项的各个 DEF，作以直接匹配扩展。

[0138] (5) 基于扩展后处理策略，根据图像库标注集信息，对扩展词集进行过滤筛选，从而获取优化后的最终扩展词集。

[0139] (6) 重复 (2) ～ (5)，获得原始查询中每个关键词项的扩展词集进行合并，将其作为与原始查询相对应的扩展后查询表达。

[0140] 5. 检索结果的评分与优化算法

[0141] 图像搜索引擎的排序基本单元为图像，其排序的基本依据为图像特征。在基于图像内容的图像检索中，图像的底层特征作为一幅图像的特征；而在基于文本的图像检索中，图像的标注信息即为图像特征。对于后者，附加用户输入的查询关键词作为图像的排序标准，排序就是将标注词序列更接近于查询关键词序列的图像排列在检索结果列表中更靠前的位置上。因此，需要检索结果的评分与优化算法，以确定作为检索结果的各幅图像相比较哪一幅对于用户查询关键词序列来说更“好”。然而，“好”的标准实际上并不存在，不同的用户即使输入同样的查询，也很有可能对同样的返回结果做出大相径庭的评价。所以，检索结果的评分与优化算法只是定义一种评分规则，通过调整参数为图像评分而进行排序，以期达到更“好”的效果。

[0142] 鉴于现有技术公开的文本检索领域中的排序处理是一项较为成熟的技术，如 [24, 25,26]，该类算法考虑更多的是查询关键词与检索文档的直接匹配，有可能造成未包含用户查询关键词但确实相关的检索文档不能被返回。本发明建立适用于基于文本的图像检索的排序策略，提供一种检索结果评分与优化算法，结果显示，对图像搜索引擎的搜索行为未

加任何干扰，对所返回的检索结果无任何影响。该算法的主要作用在于，让同一检索结果集中更“好”的结果排序更前，以便用户更容易观测到。

[0143] 图像搜索引擎的检索结果往往很多，而用户往往只会有耐心察看前面的一些结果。换句话来说，如何将更贴近用户搜索意图的检索结果放至返回结果更前面的位置上相当重要。因此，设计基于词语语义相似度的评分算法对返回的检索结果进行排序，根据查询关键词序列与图像的标注信息（即标注词序列）进行评分，从而将所返回各幅图像的得分作为排序依据。实际上，该算法的评分对象是图像的标注集，而对图像本身并没有任何认识，这是一个纯粹信任并依靠图像标注信息的评分方案。前述所讨论的词语语义相似度度量，就是为这里的排序算法能够得到“语义相近”的计算支持。

[0144] 由于用户查询关键词的输入是以随意方式进行，因此平等对待查询关键词序列中的每个词项。然而，对于图像的标注词序列，假设排在前面的词项更为值得信赖。该假设基于一个事实，即标注者倾向于首先输入图像中最突出的物体。诚然，对同一幅图像，不同的标注者具有不同的判断，并且不一定存在最突出的物体。但对于大多数图像来说，图像中的焦点物体还是非常明显。正是出于这样的考虑，评分算法中标注词的计算结果都附加权重，用于突出图像中可能的“突出”物体。由此，使用下述公式来计算图像的排序分数：

[0145] $$Score=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}w(j,m)Sim(k_i,t_j)}{\sum_{j=1}^{m}w(j,m)} \quad (1)$$

[0146] 其中，k_i 为关键词序列的第 i 个关键词；t_j 为图像标注词序列的第 j 个标注词；$Sim(k_i, t_j)$ 用于计算两个词项 k_i 与 t_j 之间的语义相似度；$w(j,m)$ 为相关权重，$w(j,m)=(m+1-j)^2$，用于突出标注序列中标注词项的前后关系；而 n 与 m 则分别是查询关键词序列与图像标注词序列所包含的词项个数。考虑图像标注词序列中的第一个标注词权重为 m^2，则相对于总权重 $\sum_{j=1}^{m}w(j,m)$，其所占比例为：

[0147] $$\frac{m^2}{\sum_{j=1}^{m}w(j,m)}=\frac{m^2}{\sum_{j=1}^{m}j^2}=\frac{6m^2}{m(m+1)(2m+1)}=\frac{6m}{(m+1)(2m+1)} \quad (2)$$

[0148] 该函数是一个递减函数，随着图像标注词序列的增大，排头词的权重影响成线性递减。如果一幅图像含有太多物体，就会使得各个物体都不会特别突出。

[0149] 需要注意的一种情形是，评分计算中存在大量重复计算。查询关键词参与所有的词语语义相似度计算，而每幅图像的标注序列都会包含至少一个查询关键词（否则该图像不会作为检索结果被返回），并且检索结果图像中也会共有大量相同标注词项，所以实际所必需的语义计算比语义计算被调用的次数少很多。通过设置一个适当大小的相似度计算结果缓存，记录下一些语义计算结果，对处理速度的提升具有很大帮助。另一方面，图像搜索引擎在处理检索结果过多而导致的分页显示时，提倡每次访问一个分页结果都重新进行搜索。如果能够使用一个缓存，将一些图像的评分结果缓存起来，那么在用户切换图像检索结果不同分页的时候，将会避免大量计算。从相似度（语义距离）结果缓存、结果文档缓存、直至底层的同义词集读取模块缓存，基于多层缓存机制对原有评分算法进行优化，可在很

大程度上节约处理时间。

[0150] 评分算法的计算方式给出的是一个精确按照语义相似度函数结果整合的分数，而实际计算中，一些近似结果同样也能完成任务。毕竟，只需要不错的排序效果，而具体的分数值并不重要。在评分计算中，每幅图像的标注词序列都具有与查询关键词相匹配的标注词，同样也具有与查询关键词语义相似度相当小（或者语义距离相当远）的标注词。标注词序列中的各个标注词具有相应的顺序权重，一个排位在后且语义相似度很小的标注词对最终检索列表的影响，是一个排位在前且与查询关键词相匹配的标注词的二十分之一还是五十分之一显然无关紧要。因此，针对这些对最终检索列表影响不大的计算结果，统一采用一种相同结果表示而非按部就班地计算，这将大大简化计算的复杂度。

[0151] 该算法的步骤用伪代码描述如下：

[0152] (1) 获得输入：原始查询关键词序列。

[0153] (2) 使用查询扩展函数得到原始查询关键词序列的扩展词集。

[0154] (3) 为每一对查询关键词项与其扩展词项计算语义相似度，并将结果全部存于缓存之中。

[0155] (4) 基于图像搜索引擎，获取与原始查询相对应的检索结果。

[0156] (5) 为检索结果中的每幅图像计算评分。如果能够从缓存中获取相关信息，则利用现成的已有结果；否则，就当作语义计算结果很大（即语义距离很远），使用一个统一的常量结果，而不再进行语义计算。

[0157] (6) 根据检索结果中每幅图像的评分，对各幅图像进行重新排序，以返回最终的检索列表。其中，在步骤 (5) 的评分过程中，所有的语义计算将作为预处理，而实际评分时无需任何语义计算，则总的语义计算次数将比优化前的评分算法降低一个数量级，以此提高处理效率。

[0158] 实施例 2 应用实例

[0159] 附图 2 是通过一个具体的实例演示上述算法流程框架的具体步骤，通过给出各算法模块的中间输出以及该框架的最终检索结果，给人以直观的理解。

[0160] 标号 (1) 与 (2) 分别为用户输入的原始英语查询和汉语查询；标号 (3) 与 (4) 分别为利用英语义类词典 WordNet 和汉语义类词典 HowNet，采用“基于扩展规则的查询扩展词选择与优化算法”所获取的与原始英语查询和汉语查询相对应的扩展词集；标号 (5) 与 (6) 分别为基于原始英语查询和汉语查询的扩展词集，利用图像搜索引擎所获取的相应检索结果；标号 (7) 与 (8) 分别为基于原始英语查询和汉语查询的初始检索列表，利用融合“基于 WordNet 的英语词语语义相似度度量算法”和“基于 HowNet 的汉语词语语义相似度度量算法”的“检索结果的评分与优化算法”所获取的最终检索结果。

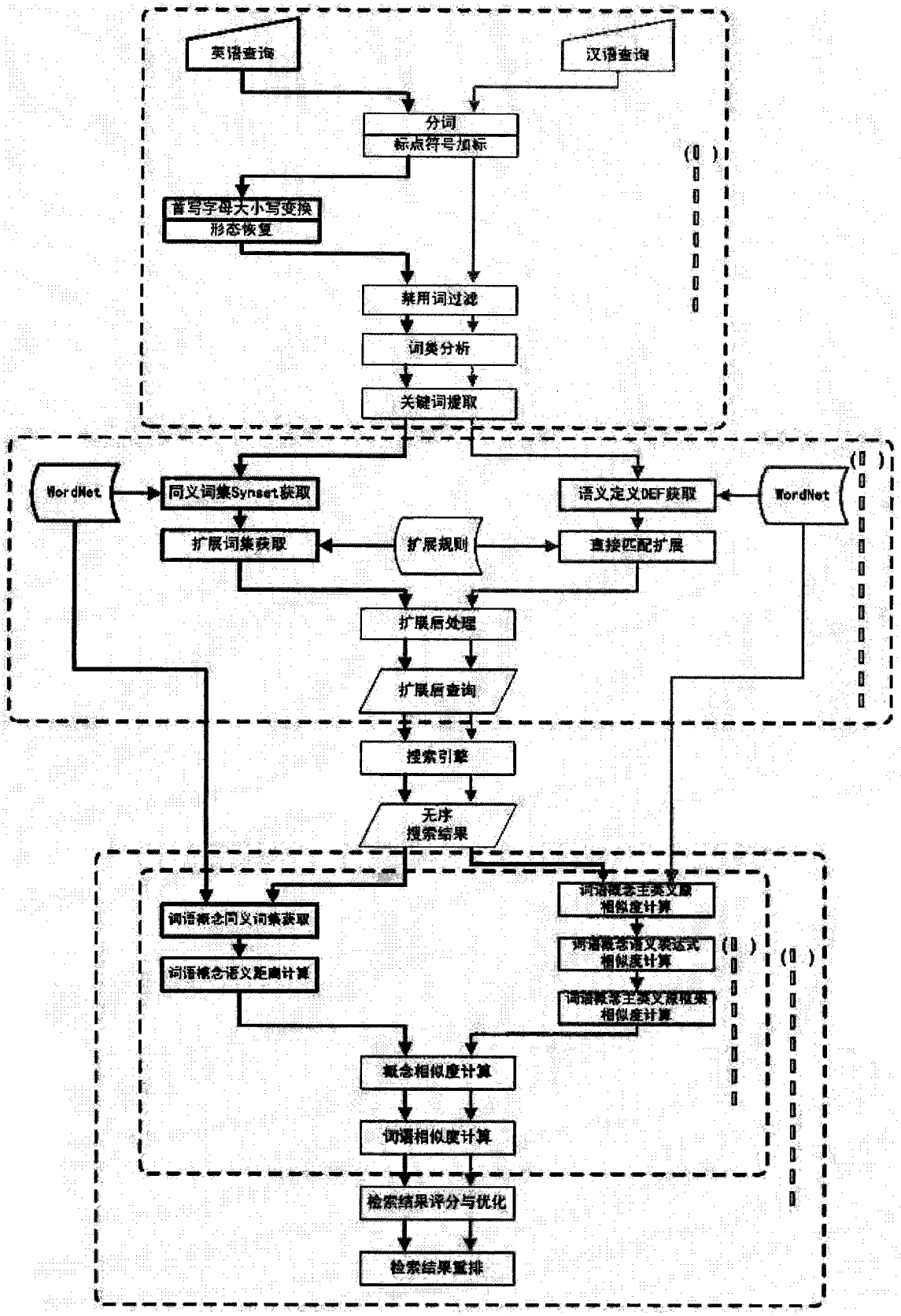
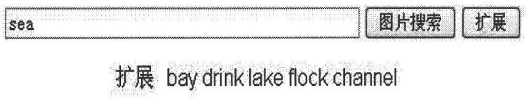
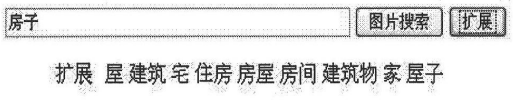

英语查询
汉语查询
分词
标点符号加标
首写字母大小写变换
形态恢复
禁用词过滤
词类分析
关键词提取
WordNet
同义词集Synset获取
语义定义DEF获取
WordNet
扩展词集获取
扩展规则
直接匹配扩展
扩展后处理
扩展后查询
搜索引擎
无序
搜索结果
词语概念同义词集获取
词语概念语义距离计算
词语概念主类义原
相似度计算
词语概念语义表达式
相似度计算
词语概念主类义原框架
相似度计算
概念相似度计算
词语相似度计算
检索结果评分与优化
检索结果重排

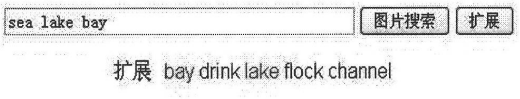
图 1


sea
图片搜索
扩展
扩展 bay drink lake flock channel

(1)


房子
图片搜索
扩展
扩展 屋 建筑 宅 住房 房屋 房间 建筑物 家 屋子

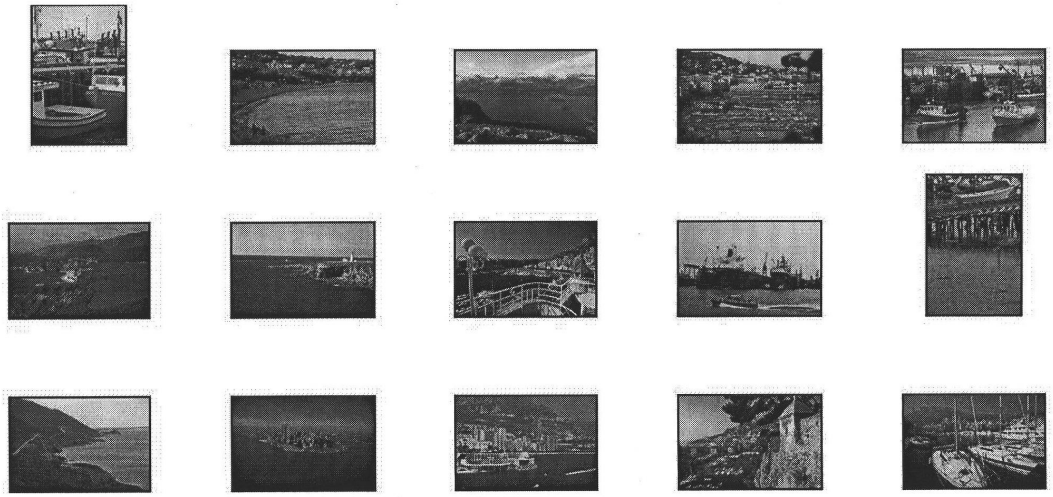
(2)


sea lake bay
图片搜索
扩展
扩展 bay drink lake flock channel

(3)

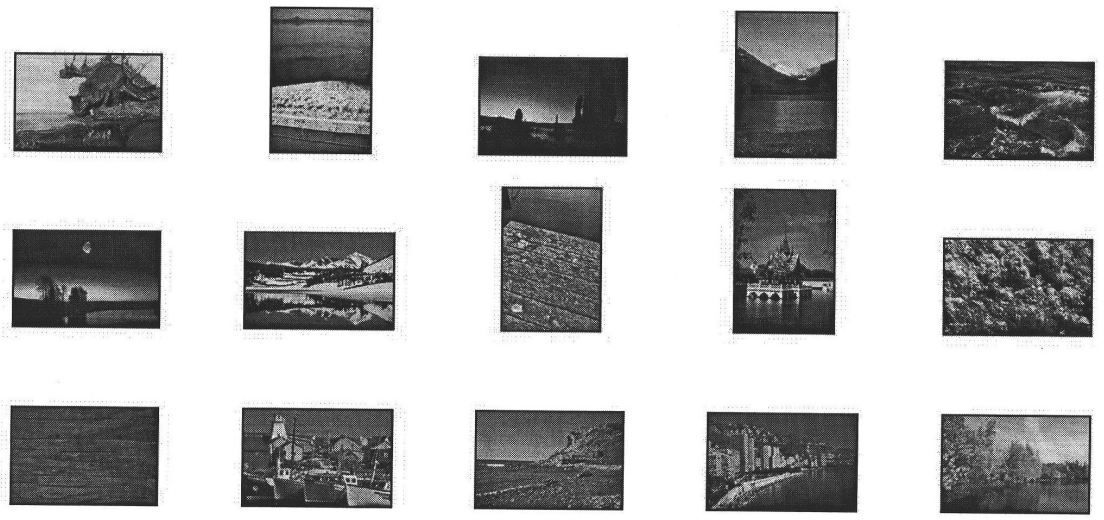

房子 建筑 房屋
图片搜索
扩展
扩展 屋 建筑 宅 住房 房屋 房间 建筑物 家 屋子

(4)

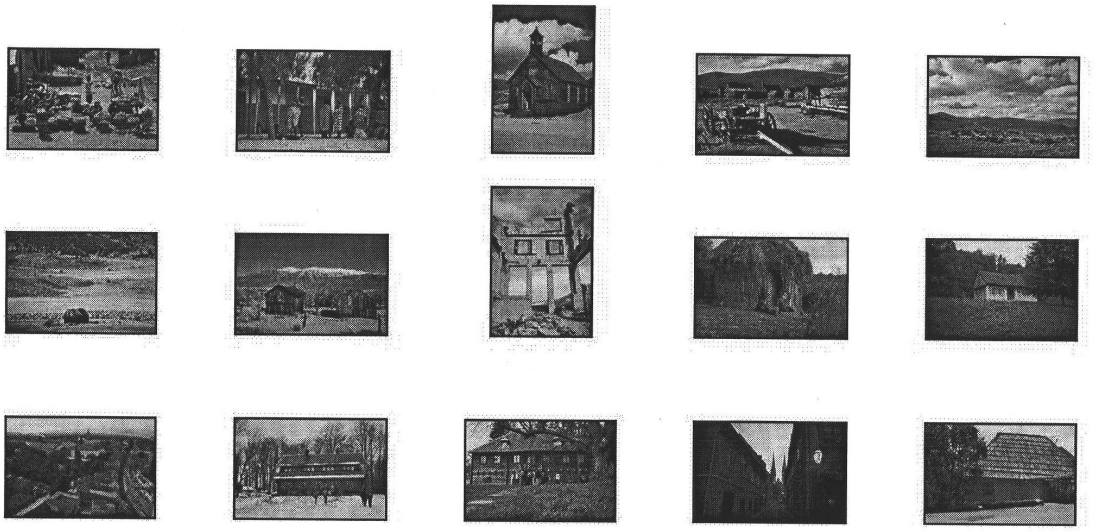


(5)

(6)



(7)



(8)

图 2