基于 XML 文本片段的图像检索实现与评价*

陆 伟 张 宓 刘 丹

摘 要 为了研究文本片段与图像位置关系对图像检索结果造成的影响,本文采用同在域范围和 RK 两种层级体系来表示文本与图像的关系,并对两种体系的不同层级分别进行检索评价。实验结果表明,就整体而言,层级越靠近图像,检索效果越好;越靠近图像的文本,越能描述图像的语义信息。RK 层级能更细致地表达文本片段与图像的位置关系,随着层级以图像实体为中心向外扩展,检索效果整体呈下降趋势。表 1。图 3。参考文献 11。

关键词 XML 检索 图像检索 TBIR
分类号 G354.4

ABSTRACT To analyze the influence of the position relation of text fragment and image on the retrieval efficiency, the field level they are both in and Region Knowledge are respectively used to present the different relations in this paper, and retrievals and evaluations are proposed on different levels. The results of these experiments indicate that closer distance between text fragment and image can bring more accurate retrieval result, and when presenting the relation of text fragment and image by RK, the direct operation on field name is avoided and the presentation of the position relation is more delicate. Among all levels centered on an image entity, the retrieval accuracy are declining from inside to outside. 1 tab. 3 figs. 11 refs.

KEY WORDS XML retrieval. Image retrieval. TBIR.
CLASS NUMBER G354.4

1 图像检索概述和研究现状

1.1 图像检索概述

随着网上图像资源的不断丰富,人们对图像检索的需求日益增多,因而如何更有效地识别并检索出相关图像成为当前信息检索研究的一个热点。目前关于图像检索的实现方法主要有两种,即基于文本的图像检索(Text-Based Image Retrieval,TBIR)和基于内容的图像检索(Content-Based Image Retrieval,CBIR)。其中,TBIR 基于全文检索技术,根据图像周围文本信息,对文档中的图像进行索引及查询,通过文本匹配检索出所需的图像;CBIR 则通过提取图像的视觉特征,如颜色、纹理、形状、空间等,对其进行索引,查询时根据这些特征的匹配情况,返回符合要求的图像结果。

当前,TBIR 是图像检索实践中应用最广泛的方法。然而,到底采用何种文本信息(是图像所在的文档、章节、段落,还是图像标题等)实现基于文本的图像检索仍然需要进一步的实验研究。可扩展标记语言(eXtensible Markup Language,XML)的出现,为系统地研究该问题提供了一条可行的途径。作为描述网络数据内容和结构的半结构化语言,XML 的结构语义标签能赋予内容清晰的逻辑层次,进而辅助实现文本片段的检索。由于 XML 文档中可以包含各种不同数据形式的多媒体信息,如音乐、图像等,因此,可以利用 XML 文本片段与多媒体信息的位置关联关系,实现基于文本的多媒体信息检索。

* 本文为国家社科基金项目"基于 XML 的多媒体信息检索模型及其实现研究"(编号:06CTQ006)研究成果之一。

1.2 研究现状

1.2.1 XML 图像检索

目前,基于 XML 的图像检索方法主要包括 MMFragments 和 MMImages。MMFragments 是根据查询需求检索 XML 文档的多媒体片段,其返回结果为包含文本和图像的 XML 文档片段;MMImages 是根据查询需求检索相关图像,其返回结果为 XML 文档中的具体图像。国外一些学者在这方面开展了深入研究,同时也出现了专门致力于 XML 检索研究的国际会议,如 INEX(Initiative for the Evaluation of XML Retrieval)等[1]。

MMFragments 实际上是对 XML 文本片段的检索,在查询时,输入的是文本信息。而与一般图像检索不同的是,其返回结果不单纯是图像,而是包括图像和文本信息的 XML 多媒体片段。关于 XML 文本片段检索的研究现状请参见文献[2]。

MMImages 方法查询时可同时输入文本和图像信息:对于文本信息,首先进行 XML 文本片段检索,得到相关节点的路径,再利用文档结构,选择同一文档片段中相关的图像作为结果返回;对于图像信息,通过一般的 CBIR 方法检索出相似的图像。自 2005 年以来,INEX 将基于 XML 的多媒体检索纳为其子任务,一些专家学者以此为平台进行了基于文本和基于内容的图像检索归并研究,如 D. N. F. Awang Iskandar 等[3]在文本检索系统和图像检索系统中分别进行检索实验后,通过线性方法归并二者结果,并变换归并系数进行评价比较,得到线性算法中的最佳归并点;D. Tjondronegoro 等[4]在二者归并过程中,以 Images、Features 和 AdhocXML 为主体建立索引,分别记录图像的名称位置、内容特征、XML 路径,以此进行图像检索;Yu Suzuki 等[5]采用文本检索中的 TF-IDF 方法对图像特征进行索引,开发 CBIR 系统,同时结合文本检索结果,构造出一个完整的图像检索模型。

1.2.2 基于 XML 文本片段的图像检索

从 INEX 2005 - 2007 年提交的论文来看,对于 MMImages 方法,研究者较多地讨论了 CBIR 方面,而对于文本查询,仅仅采用单纯的 XML 文本检索方法进行类似 MMFragments 的片段检索,并未明确讨论文本片段元素的选择问题。因此,本文将不考虑基于内容的图像检索问题,而将研究焦点集中在 MMImages 检索中利用不同层次的 XML 文本片段进行图像检索效果的考察上。

为了将检索出的文本片段与图像联系起来,就需要考察文本与图像之间的位置关系,研究不同关系对检索结果造成的影响。对此,伦敦 Queen Mary 大学的 Zhigang Kong 等[6-7]提出了域知识(Region Knowledge,RK)的概念,其以一个具体的图像实体元素为中心,将 XML 文档中的所有文本内容看作是对该图像实体的描述(Knowledge),然后根据这些内容所在元素与图像实体的逻辑关系,将不同文本划分到不同区域(Region)层级中。因此,对于文档中的所有文本,都可以采用区域层级来描述它与图像实体的位置关系。

Mouna Torjmen 等人[8-9]采用类似 RK 的原理,在实验中选取与图像实体最近的四个节点,即图像实体本身及其父节点、子节点、兄弟节点。检索时,对于某一图像,采用 XML 文本检索系统 XFIRM,分别计算其四个相关节点的文本片段权值,并按一定系数将它们线性相加,作为该图像的最终得分。但该实验所涉及的相关节点并不完整,因此检索范围较小。

基于以上理论及成果,本文尝试进一步研究文本片段与图像位置关系对检索结果造成的影响,在以二者关系划分的不同层级上,分别进行检索实验,并评价分析实验结果。

2 研究思路与方法

2.1 检索实现的思路

本实验针对文本检索主题,先实现 XML 文本片段的检索,再根据文本与图像关系层级,在文本检索结果所涉及到的文档中,结合图像路径索引,进行图像检索,最后将文本权重直接作为相关图像的得分,返回图像结果。整个检索过程的体系结构如图 1 所示。

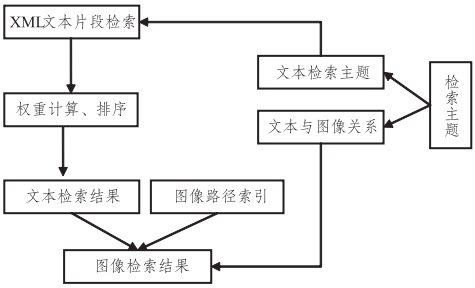

图 1　检索过程体系结构图

本文对于文本片段得分的计算,仍然采用概率公式 BM25[10] 的扩展,如下公式所示[11]:

$$w_j(e,d,C)=\frac{(k_1+1)tf_{e,j}}{k_1(1-b)+b\dfrac{el_e}{avel}+tf_{e,j}}\log\frac{N-df_j+0.5}{df_j+0.5}$$

其中,C 表示文档集,d 表示一个文档,e 表示文档 d 中的一个元素片段,k_1 和 b 是调和参数,N 是文档数,$tf_{e,j}$ 是词 j 在片段 e 中的频次,el_e 是片段 e 的长度,$avel$ 是文档片段的平均长度,df_j 是文档集中出现了词 j 的文档数。

2.2　XML 文本与图像路径关系表示方法

为了研究文本片段与图像位置关系对检索结果造成的影响,需要引进一定的方法体系来表示文本片段与图像的路径关系级别。本研究采用了两种方法定义二者关系,即同在域范围关系定义和 RK 层级关系定义。实验中,针对特定的检索词,系统在两种体系的各层级上分别进行检索,并对比结果进行分析评价。

2.2.1　基于同在域范围关系定义

XML 文档是树状结构的,如果把树状结构的每一个节点看作一个域,则 XML 文档由多个域嵌套组成。域范围由小到大可以分为多个层级,如以本文采用的 INEX – WIKI 数据集中的文档为例,包括 < body > < /body > (表示文章正文部分)、< section > < /section > (表示文章某一章节)、< p > < /p > (表示文章某一段落)和 < figure > < /figure > (表示一个图像)等。因此,在一个固定的 XML 文档集中,从大到小选出几个有代表性的域,检索时,规定出域的层级,则返回图像与文本片段应同在该层级的域范围内。如本实验关注三个域,即 < body > < /body > 、< section > < /section > 和 < figure > < /figure > ,根据这三个域所包含的范围不同,图像检索可以分成三个级别。查询时,选择某个级别的域,则检索出的图像应与文本检索得到的某条路径在该级别的同一域内。因而对于用户来说,查询时输入的关键词描述相应所选域范围内的所有图像,即指定 < body > 域时,关键词与图像需在同一篇文档内,一旦在某文档中查询到关键词,则该文档 < body > 域包含的所有图像都作为结果返回;指定 < section > 域时,关键词与图像需在文档中的同一章节内,在文档某章节查找到关键词时,则该章节包含的所有图像都作为结果返回;指定 < figure > 域时,关键词与图像需在同一个图像元素内,关键词出现在图像标题或图像描述中。

2.2.2　基于 RK 层级关系定义

Zhigang Kong 等在实验中,以某个图像实体为中心,将 XML 文档中的全部文本作为描述图像的信息。根据文本所在节点与图像实体所处位置的树状关系,将文本信息划分为不同级别的 RK,则在一篇最大深度为 N 的 XML 文档中,包含 N – 1 个层级的 RK(如图 2 所示):

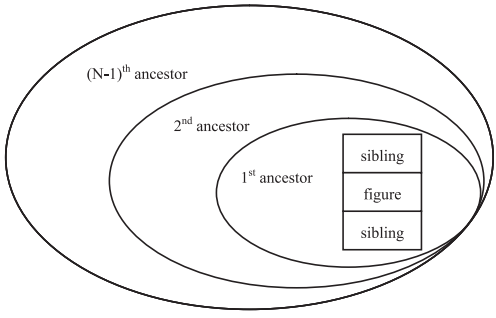

图 2　RK 层级关系结构

- Self Level RK:文本信息与图像在同一个元素内,该元素整体作为描述图像信息的一个实体。在本文采用的数据集中,这个元素即 < figure > < /figure > 。因此,所有在 < figure > < /figure > 内的文本信息(包括该域内其他元素如 < title > < /title > 之间的文本)都称作这一图像的 Self Level RK;

- Sibling Level RK:文本信息所在的元素,与图像实体元素之间为兄弟关系,并且两个元素的位置相邻;
- 1st ancestor level RK:文本信息所在的元素,是图像实体元素的父节点元素;
- ……
- (N-1)th ancestor level RK:文本信息所在的元素,是(N-2)th ancestor level RK所在元素的父节点元素。

Zhigang Kong 等在定义 RK 时,对于任意 nth ancestor level RK,其文本信息不包括已用作低级别 RK 的文本节点,而本实验意在探讨采用不同层级时的检索效果,因而低级别 RK 的文本信息包含在高级别 RK 中,也就是说,图 2 中除 sibling 层级与 figure 平级外,其他层级以 figure 为中心,不断向外扩展。本实验考虑图像实体周围的 6 类 RK——self、sibling、1st ancestor、2nd ancestor、3rd ancestor、4th ancestor,它们分别表示文本所在节点是图像实体的本身、兄弟、父节点、祖节点、3 级祖先、4 级祖先。在这些层次上,分别进行检索,并对比分析结果。

3 检索实验与评价方法

3.1 实验数据集

如上文所述,本研究采用 INEX 提供的WIKI数据集[1],该数据集以维基百科中的英文数据为基础,包含多个集合,本实验采用其中的 Wikipedia Ad Hoc XML collection、Wikipedia image collection 和 Wikipedia image XML collection 三个集合。其中,Wikipedia Ad Hoc XML collection 是 INEX 用于文本检索的 XML 数据集,文档中每个图像都以 < image > 标签用于识别;Wikipedia image collection 包含了 Wikipedia Ad Hoc XML collection 中提到的所有图像;Wikipedia image XML collection 由各图像的元数据文档组成,每个文档均包含了一个图像以及该图像的相关信息。

对于 INEX - WIKI 数据集的图像检索,INEX官方给出 20 个查询主题。官方检索主题示例中同时包含了 CO(Content Only,基于内容的)和 CAS(Content and Structure,内容 + 结构)的查询主题。但是,由于实验数据集较小,这里仅仅采用简单的词汇进行 CO 的查询。本实验中构造了 20 个检索主题:Golden Gate Bridge、Great Wall、Mickey Mouse、nebula、Pepsi、Taekwondo、gymnastics、trumpet、piano、Wal-Mart、world trade center、Arc De Triomphe、Castle、Flower、Knot、National flag、George W Bush、Saddam、Hepburn 和 Jackie Chan。

3.2 基于 WHU - XML 的实验系统设计

WHU - XML 是由陆伟等[11]在改造Okapi全文检索系统的基础上开发的 XML 半结构化信息检索系统,它不仅支持文档级 XML 检索,同时支持元素级(片段级)的 XML 索引与检索。

为支持图像检索,笔者对 WHU - XML 系统进行了改造,对图像和文本路径之间的关系建立索引,以支持图像和文本片段位置关系的识别,进而利用文本片段实现图像的检索。根据上文 2.2 中表示文本片段与图像位置关系的两种方法,笔者设计出相应的两种检索模式,即检索模式 1 和检索模式 2,并开发了相应的前台图像检索接口界面。

检索模式 1 采用同在域范围定义关系层级。该模式根据文本与图像所同在域的范围,选取域 < body > </body > 、< section > </section > 和 < figure > </figure > 表示所有关系。其具体实现思路是:依次读取文本检索结果中的文本片段,如果该片段处在所指定的域内,则将其权值作为该域下所有图像的权值,并将图像作为结果返回。

检索模式 2 采用 RK 定义关系层级。该模式采用 RK 方法,选取其中的层级 self、sibling、1st ancestor、2nd ancestor、3rd ancestor 和 4th ancestor 表示文本与图像关系。其实现思路是:依次读取文本检索结果中的文本片段,将其与同一文档中的所有图像一一比较,计算二者路径所在的层级差,筛选出符合要求的图像。

检索时,先选择模式并输入层级,系统在众多图像中,根据输入的层级关系,选出相关的图像作为结果返回,并将路径权值作为图像权值,

对图像进行排序输出。当同一文档下的同一图像出现多次时,则只选择权值最高的结果,而过滤掉其余重复的结果。

3.3 实验过程

实验时,首先对数据集进行索引。索引过程包括文本索引和图像索引两方面。其中,文本索引是原有 WHU - XML 系统根据其索引机制,对数据集中各文档进行预处理,并在此基础上进行内容索引和结构索引,生成一系列索引文件。图像索引则是通过建立图像数据库,写入数据集中所有图像的编号、名称及路径位置等信息,以便查找与匹配。

检索时,对于每个查询主题,实验中都采用了两种检索模式,针对不同层级进行检索及统计。在模式 1 中,针对其 3 个不同的域(body、section、figure)分别进行检索。在模式 2 中,针对 6 种层级关系(self、sibling、1st ancestor、2nd ancestor、3rd ancestor、4th ancestor)分别进行检索。

3.4 评价方法

查准率和查全率是信息检索效率评价的两个定量指标,不仅可以用来评价检索的准确性和全面性,也是信息检索系统评价中衡量系统检索性能的重要方面。查准率反映了检索的准确性,查全率则反映了检索的全面性。本次实验着重关注查准率,对每个主题在各层级上的检索分别使用 Precision、MAP 及 P@10 等三项指标进行评价。

在本实验中,Precision(检准率)是指检索结果中相关的图像数量与所有被检索出来的图像数量的比例。Precision 是早期信息检索最常用的评价方法之一,本实验重在研究层级对检索结果的影响,因此需考察各层级上的整体结果,所以 Precision 仍是很重要的评价指标。

单个主题的 MAP(Mean average precision,平均准确率)是每个相关图像被检出后的准确率的平均值。主题集合的 MAP 是每个主题 MAP 的平均值。MAP 是反映系统在全部相关文档上检索性能的单值指标。系统检索出来的相关文档越靠前,MAP 就越高。

P@10(Precision at 10)是指检索出前 10 个图像的准确率。由于用户对靠前的检索结果更为关心,因此以该指标考察本实验系统中第一页返回结果的准确率。

4 实验结果评价分析

4.1 实验结果评价

本实验采用上文提到的 20 个检索主题,针对两种检索模式的各层级分别进行检索。对于各评价指标,本文取各主题检索结果的平均值,得到表 1。

表 1 实验结果

检索模式	层级	Precision	MAP	P@10
模式 1	body	0.23	0.61	0.72
	section	0.29	0.56	0.33
	figure	0.79	0.92	0.79
模式 2	self	0.79	0.92	0.79
	sibling	0.25	0.57	0.52
	1st ancestor	0.28	0.62	0.59
	2nd ancestor	0.23	0.61	0.47
	3rd ancestor	0.18	0.45	0.28
	4th ancestor	0.16	0.34	0.38

4.2 实验结果分析

通过表 1 可以看出,对于检索模式 1,从 body、section 到 figure,随着图像与文本片段同在域范围不断缩小,Precision 不断上升;在 body 和 figure 层级下,P@10 值很高,说明靠前的图像检准率高;当取层级 section 时,各指标值均较低,这是否是因为 section 标签存在于文档中不同层级,还需进一步实验。

模式 2 中除 sibling 与 self 平级外,其余各层级均以 self 为中心,不断向外扩展。分析这些层级的检索结果,可得到图 3。从图中可看出,随着层级向外扩展,各指标值总体呈下降趋势,表明层级越靠近图像,检索效果越好,也说明越靠近图像的文本与图像的语义关系越紧密。

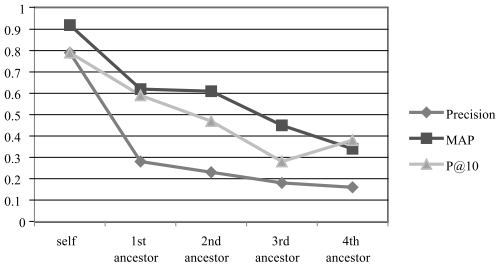

图 3　模式 2 下检索结果评价各指标曲线图

对比两种检索模式,可以看出模式 1 的 figure 层级相当于模式 2 的 self 层级,表示文本内容与图像同在一个图像实体内,该层级的各项指标均达到最大值。

由于 XML 仅仅是半结构化的,因此标签的应用较为灵活。模式 1 中直接以域名称表示范围,会造成某些文档中特殊的域得不到检索;模式 2 采用 RK 层级表示文本片段与图像的位置关系,避免了对域名称的直接操作,能够查找到更多信息,同时二者关系也表达得更加透彻。

5　总结与展望

本文针对 XML 文本片段与图像位置关系对图像检索结果造成的影响,进行了实验及初步分析。研究结果表明,整体而言,越靠近图像的文本,越能描述图像的语义信息。此外,RK 层级能更细致地表达文本片段与图像的位置关系,随着层级以图像实体为中心向外扩展,检索效果整体呈下降趋势。然而,在本研究中,对于重复图像的处理、多数据集及更大数据集的评测、文本片段加权对图像检索效果的影响等,尚未进行深入考虑,有待在以后的研究中进一步深化。

参考文献:

[1] INEX 2007[OL]. [2008-07-01]. http://inex.is.informatik.uni-duisburg.de/2007/.

[2] 陆伟. 元素级 XML 检索模型构建的关键问题与解决方案研究[J]. 中国图书馆学报, 2007(6): 60 - 63.

[3] D. N. F. Awang Iskandar, Jovan Pehcevski, James A. Thom, S. M. M. Tahaghoghi. Combining Image and Structured Text Retrieval[C]. Advances in XML Information Retrieval and Evaluation, Springer Berlin/Heidelberg, 2006: 525 - 539.

[4] C. Lau, D. Tjondronegoro, J. Zhang, S. Geva. Fusing Visual and Textual Retrieval Techniques to Effectively Search Large Collections of Wikipedia Images[C]. Comparative Evaluation of XML Information Retrieval Systems, Springer Berlin/Heidelberg, 2007: 345 - 357.

[5] Yu Suzuki, Masahiro Mitsukawa, Kenji Hatano, Toshiyuki Shimizu , Jun Miyazaki, Hiroko Kinutani. An XML Fragment Retrieval Method with Image and Text using Textual Information Retrieval Techniques[C]. INEX 2007 Workshop Pre-Proceedings, 2007: 433 - 435.

[6] Zhigang Kong, Mounia Lalmas. XML Multimedia Retrieval[C]. SPIRE 2005: 218 - 223.

[7] Zhigang Kong, Mounia Lalmas. Using XML Logical Structure to Retrieve (Multimedia) Objects[C]. ECDL 2007: 100 - 111.

[8] Lobna Hlaoua, Mouna Torjmen, Karen Pinel-Sauvagnat, Mohand Boughanem. XFIRM at INEX 2006. Ad-Hoc, Relevance Feedback and Multi-Media Tracks[C]. Comparative Evaluation of XML Information Retrieval Systems, Springer Berlin/Heidelberg, 2007: 373 - 386.

[9] Mouna Torjmen, Karen Pinel-Sauvagnat, Mohand Boughanem. MM-XFIRM at INEX Multimedia track 2007[C]. INEX 2007 Workshop Pre-Proceedings, 2007: 423 - 432.

[10] Stephen Robertson, Hugo Zaragoza, Michael Taylor. Simple BM25 Extension to Multiple Weighted Fields[C]. CIKM 2004: 42 - 49.

[11] 陆伟, 夏立新. 基于 OKAPI 的 XML 信息检索实现研究[J]. 中国图书馆学报, 2006(4): 60 - 64.

陆　伟　武汉大学信息资源研究中心研究员,博士,副教授。通讯地址:武汉大学信息管理学院。邮编 430072。

张　宏　武汉大学信息管理学院 2004 级本科生。通讯地址同上。

刘　丹　武汉大学信息管理学院 2007 级硕士研究生。通讯地址同上。

(收稿日期:2008-07-05;修回日期:2008-08-06)