浙江大学

本科生课程论文

题　　目　旅游搜索引擎的推荐系统和知识图谱构建研究

姓　　名　薛　伟

学　　号　3150101324

专　　业　软件工程

指导老师　邢卫 邵建 金波

所在学院　计算机科学与技术学院

2018 年 7 月 24 日

旅游搜索引擎的推荐系统和知识图谱构建研究

软工 1503

3150101324

薛　伟

摘　要：许多游客总是在网上搜索旅游景点相关信息，以获得他们正在访问的地点或计划下次旅行的更多信息。 在我们对搜索引擎的研究中这项研究中，引入了旅游景点推荐系统，提供特定旅游景点的相关知识，并根据特定旅游景点和之间的语义相关性为其他相关的地方提供建议，引入了关于推荐系统的实现和计算方法，以及知识图谱的构建，为完善的推荐系统提供了研究的思路。

关键词：语义关联度;推荐系统;知识图谱;用户兴趣;

1　引 言

我们在本次项目实训课程中，完成的是一个旅游线路攻略的垂直搜索引擎，通过对目前旅游网站、其他媒体资源内容的爬取，建立自己的一套知识体系，最终完成搜索引擎的构建。在项目中，非常重要的一项内容，就是知识体系的构建，在这里我们采用了知识图谱的方法，由于是初涉关于这方面的内容，所以我们对于知识图谱做了一些研究和思考，尽管最后在实现时较为简单，不过在过程中还是有许多想法和收获，所以在此论文中详细阐述。

要知道，与特定旅游景点相关的信息始终分布在不同的 Web 源中，所以我们必须要整合不同来源的各种信息，最终整合成统一的信息提供给搜索引擎来搜索和呈现。而这样旅游信息搜索引擎的最重要的一点是，搜索结果的呈现顺序，或者说关联度/推荐度的排序，所以我们建议，当获得某些先前的访问信息时，可以基于统一的搜索引擎信息自动进行有趣的旅游景点的建议。语义相关性在开发这种主动推荐算法中起着重要作用。

在本文的内容中，主要探讨了两种旅游景点积极推荐策略的尝试，和旅游景点知识图谱的构建。接着研究了旅游景点推荐的语义相关度计算算法，并就我们的实现作出进一步的讨论。

2 基于用户兴趣和语义相关度的主动推荐策略

在给定之前特定用户浏览信息的情况下，可以退出两种基于语义关联度的推荐方式来满足不同用户的相应偏好。在此使用一些较为正式的数学符号来表示这两种推荐策略。

策略一：正相关推荐。如果用户（使用 v 来表示）对某个旅游景点（使用 t 来表示）感兴趣，那么就会有一系列备选景点（记为 $T=\{t_1, t_2, \cdots, t_n\}$，一个有序的列表 T’会被推荐给用户。推荐的顺序正相关于备选列表 T 内的各个景点 t 的语义相关度。所以对于每一个备选景点，可以得出如下的规则：

$$K_r I_v t \wedge K_r(t \sim t') \rightarrow K_r I_v t \wedge K_r I_v t'$$

其中 r 表示推荐系统，K 是一个认知常数，I 是兴趣常数，t’是列表 T 内的一个元素。t～t’表明两个景点在语义上是相关的。所以这个规则表明如果 r 知道 v 之前对 t 有兴趣（访问过 t），而 t 在语义上相关 t’，那么 r 会将 t’推荐给 v。对于喜欢这个策略的用户，他们喜欢的旅游景点应该在某种程度上相似。

策略二：负相关推荐。如果用户（使用 v 来表示）对某个旅游景点（使用 t 来表示）感兴趣，那么就会有一系列备选景点（记为 $T=\{t_1, t_2, \cdots, t_n\}$，一个有序的列表 T*会被推荐给用户。而推荐顺序负相关于备选列表 T 内各个景点 t 的语义相关度。使用这种策略，推荐系统假定 v 喜欢之前没有体验过的景点。所以可以得出以下规则：

$$K_r I_v t \wedge K_r \dashv (t \sim t') \rightarrow K_r I_v t \wedge K_r I_v t'$$

其中 $\dashv (t \sim t')$ 表示两个景点在语义上不相似。对于喜欢这种策略的用户来说，与之前的访问相比，旅游景点的新颖性非常重要。

3 旅游景点（路线）的知识图谱

知识图谱或者说知识库在各种智能系统中发挥着核心的作用。那么旅游景点知识图谱也是搜索引擎推荐系统的核心。我们所设计的旅游景点的知识图谱遵循着一般的语义知识图谱的设计原则。它从各种资源中提取和整合多模态知识（各种旅游网站中），然后将不同的实体进行标注、链接，形成统一的一个完善的知识图谱。

其中，不同旅游景点的名称是收集和整合信息以建立旅游景点知识图谱的关键。在我们的研究中，考虑采用两种策略来收集这些名称。首先，选择爬虫到的上层概念的一些名词，主要是基于旅游网站的一种详细陈列。其次，选择“中国旅游景点评级”中列出的名称（从 A 级到 AAAAA 级）。有 12275 个旅游景点，最终包含在旅游景点知识图谱中。

从不同来源提取几种类型的信息以构建知识图谱。在百科全书网页中，我们提取以下内容：

（1）信息框中的内容（它们被提取并存储为关于特定旅游景点的三重形式声明性知识）；

（2）每页开头的介绍部分（这些段落用于快速介绍特定的旅游景点。它们作为文本资源存储，并通过二元关系与特定的旅游景点实体相连）；

（3）提取“开放分类”和“相关项目”部分下的词语术语（“开放分类”下的术语是旅游景点的上层概念，而“相关项目”下的词语被认为是相关资源，如作为类似的旅游景点，这两种资源对于定量计算旅游景点之间的语义相关性至关重要;

（4）嵌入旅游景点页面的图像（该资源将为访问者提供额外的可视化信息）。从图像搜索引擎，我们为每个旅游景点提取至少 10 张图片。所有资源都存储为知识库中的内容（如表 1 所示），以便可以在语义方法中访问和使用它们。

表 1 一个景点的景点知识内容

内容类型	标题	标签	内容
信息性内容	国家体育馆	其他名字	鸟巢
	国家体育馆	高度	69 米

图片	国家体育馆	图片	http://img1.gtimg.com/news/pics/22304/22304186.jpg
介绍部分	国家体育馆	介绍	“...”

知识库内的知识三元组并不是孤立的。 相反，它们中的大多数彼此连接以形成大型知识网络。 图 1 显示了中国 297 个旅游景点的知识网络。 如图所示，网络中有几个关键节点，其节点度远大于其余节点。 我们发现他们中的大多数是各种旅游景点的上层概念。它们是连接不同旅游景点和其他相关知识的枢纽。它们也是构建知识图谱并计算不同旅游景点之间的语义相关性以提出建议的关键。

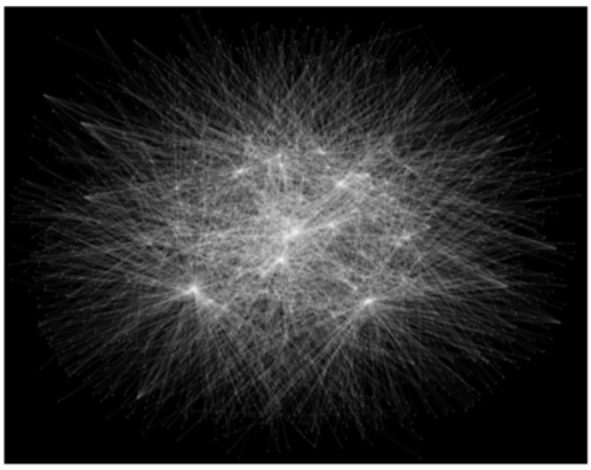

图 1 可视化的知识网络图

4 旅游景点（路线）的语义相关性计算

在第 2 节中，曾讨论到两种积极推荐旅游景点的策略。它们都依赖于先前访问地点和候选地点的语义相关性。在本节中，将对具体的计算方法作出讨论，用于根据旅游景点知识库中的知识来计算实体语义相关性。

根据之前的标记设定 t 是以前访问过的旅游景点或者被查找的旅游景点，而 $T=\{t_1, t_2, \cdots, t_n\}$ 是候选旅游景点列表。两个因素是所提算法的关键。首先，如第 3 节所述，不同旅游景点的各种上层概念将它们连接在一起。因此，共同的上层概念从一个角度表明不同旅游景点的语义相关性（例如，“故宫博物馆”和“明

十三陵”分享上层概念“世界文化遗产”，“文化”，“旅游”等）。在之前完成的知识图谱中有几个单词术语作为“相关项目”（例如，“颐和园”拥有几个“相关项目”，包括“故宫博物院”，“明十三陵“和”天坛“等）。这些相关项目由百科全书用户手动标记（有时存在协商过程，以便大多数用户同意列出的项目）。因此，它们共同反映了各种用户对其是否相关的看法。但只有这些信息，仍然难以定量衡量这些相关项目的相关性。因此，第二因素用作附加源，并且这两个因素需要组合在一起用于语义相关性计算。在上层算法中，如果特定候选旅游景点 t_n 被列为 t 的“相关项”，则基于共同上层概念计算的语义相关性将通过权重 ω 加强。具体的计算过程较为简单，但是各种权重参数 ω 及 K/I 为直接通过文献资料获取，对其获得过程不是很清楚，所以在此不做过多的讨论。

5 旅游景点（路线）推荐系统

我们最终实现/想要实现的一套系统，是在旅游攻略路线信息搜索引擎中，加入可行而优化的推荐系统。该系统的用户输入是旅游景点的名称（它可以是以前访问的地方或访问者感兴趣的地方）。该系统的输出是:

（1）相关的文本信息和关于被查询的旅游吸引力的图像;

（2）对下一次旅行可能有趣的旅游景点的建议。

推荐系统为访问者提供的旅游景点信息包括以下两种类型：

（1）不经常更新的信息（来自信息框的声明性知识，上层概念，相关项目和基于 Web 的百科全书的介绍部分。图 3 示出了呈现旅游景点的陈述性知识的功能;

（2）频繁更新的信息：与特定旅游景点相关的新闻和微博帖子。请注意，经常更新的信息起着独特的作用。例如，当旅游景点举办特殊活动时，这种信息很可能出现在新闻网页和微博帖子中。此外，最新信息对潜在访客可能非常重要。例如，如果在这两天在新浪微博上得到长城人满为患的信息，他/她可能会重新考虑访问计划以避免过于拥挤的访问。

构建推荐系统的门户网站的信息源如图 2 所示（即百度百科全书，互动百科全书，游侠客，携程，马蜂窝，新浪微博等）。

图 2 信息爬取来源

在搜索推荐的的左栏，根据输入提供了推荐的旅游景点列表。单击特定建议时，会向用户显示此建议的原因，如表 2 所示。在查询推荐系统时，如果系统共享相同的名称，系统可能会产生几个候选旅游景点。 例如，如果查询词是“故宫博物馆”，推荐系统将分别有五个候选人，分别是北京，南京，沉阳，台北和首尔的“故宫博物院”，因为他们在知识库中拥有相同的标签。 和推荐系统无法确定访问者正在寻找哪一个。

表 4 推荐分析内容

旅游景点	故宫博物院，明皇陵
共同标签	艺术，世界文化遗产，休闲，历史，民族，宫，建筑，施工，文化，文物，旅游，旅游景点，生活，纪录片，废墟，遗产公园

在大多数情况下，可以通过考虑共同出现的上下文文字信息来处理实体消歧问题。在这项研究中，访客只有“颐和园”这个词作为推荐系统的输入。需要提出另一种策略。在这项研究中，我们假设某个特定的游客正在查找相关的旅游景点信息，他/她可能与旅游景点在同一个城市。

6 总结与展望

本文主要介绍了我在实现垂直搜索引擎的推荐系统的一些研究，在提供旅游路线搜索功能的同时实现了主动的推荐系统。本文提出的推荐策略基于语义相关度计算和知识图谱的构建。相关性计算算法利用与特定旅游景点相关的两个知识源，即本体结构和相关项目。经过个人研究表明，提出的算法效果较为优化。所提出的策略和算法集中于语义及其相关计算如何在旅游景点推荐中发挥作用，因此它们通常可被视为基于内容的过滤技术。知识图谱在推荐系统中发挥了较为重要的作用，作为算法的基础数据源和算法的实现基础。

但我目前的实现和考虑还比较简单，如果想要更好地构建知识图谱，更好地去实现推荐系统，还需要考虑和集成更多因素和方法（例如协同过滤和用户个性化）以获得更好的推荐性能。

在此，作为项目实训拓展研究的内容，确实获得了不小的收获，感谢各位组员的积极配合，感谢老师的指导和奉献。

参考文献

【1】 Zeng, Y., Huang, Z., Liu, F., Ren, X., Zhong, N.: Interest Logic and Its Application on the Web. In: Xiong, H., Lee, W.B. (eds.) KSEM 2011. LNCS (LNAI), vol. 7091, pp. 12－23. Springer, Heidelberg (2011)

【2】 Douglas, B., Lenat, C.Y.C.: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM 38(11), 32-38 (1995)

【3】 Zeng, Y., Wang, D., Zhang, T., Wang, H., Hao, H., Xu, B.: CASIA-KB: A Multi-source Chinese Semantic Knowledge Base Built from Structured and Unstructured Web Data. In: Kim, W., Ding, Y., Kim, H.-G. (eds.) JIST 2013. LNCS, vol. 8388, pp. 75-88. Springer, Heidelberg (2013)

【4】 钟青.Web 旅游文化挖掘中的实体关系抽取及知识链接系统构建[D].江西财经大学,2016.

【5】 桂林电子科技大学.一种基于知识图谱的个性化旅游路线规划方法:中国,CN201710919660.3[P].2018-02-09.

【6】 余丽,陆锋,张恒才.网络中文文本蕴含地理实体关系的无监督抽取方法[C].//第六届全国地理信息科学博士生学术论坛论文集.中国科学院地理科学与资源研究所%中国科学院地理科学与资源研究所,2014:85-87.

【7】 朱桂祥,曹杰.基于主题序列模式的旅游产品推荐引擎[J].计算机研究与发展,2018,(5):920-932. DOI:10.7544/issn1000-1239.2018.20160926.

【8】 王潇慧.基于 Web 的旅游产品推荐系统设计与研究[J].现代电子技术,2018,(10):97-99,104. DOI:10.16652/j.issn.1004?373x.2018.10.025.

【9】 孙彦鹏,古天龙,宾辰忠, 等.基于多重隐语义表示模型的旅游路线挖掘[J].模式识别与人工智能,2018,(5):462-469.
DOI:10.16451/j.cnki.issn1003-6059.201805008.

【10】 尹书华,傅城州.基于百科大数据的旅游景点推荐系统应用研究[J].旅游论坛,2017,(3):107-115. DOI:10.15962/j.cnki.tourism forum.201703035.