基于 KG-Fuzz 的语义搜索引擎测试实验报告

MF1933058 刘凡维

1.研究动机

搜索引擎可能存在错误，误导用户，给用户灌输错误的知识，有可能导致严重的后果。对于搜索引擎测试，用户只能在不了解其设计细节的情况下使用在线搜索服务。因此测试的难点在于如何构造大量的“搜索”以及对返回的结果作出评价。

本次实验旨在通过测试找到搜索引擎中那些没有返回正确答案的“搜索”，并将它们进行分类总结，以归纳搜索引擎中存在的缺陷的一般模式，提升用户使用搜索引擎时的用户体验。

2.研究背景

大多数软件测试技术都假设 oracle 是可用的，测试人员可以根据 oracle 来验证测试用例执行结果的正确性。但是，在某些情况下，oracle 不可用或使用成本太高，这种情况称为 oracle 问题，是软件测试的一个基本挑战。为了缓解 oracle 问题，开发了一种蜕变测试（Metamorphic testing）方法。与传统的测试方法不同，蜕变测试不关注对每个单独输出的验证，而是检查被测程序的多个执行的输入和输出之间的关系。

同时，以三元组形式存储数据的知识图谱给 query 的简单构造提供了可能，并有助于缓解 oracle 问题。对于 query 的构造，可能采用以下方法：

基于一个三元组<E1,R,E2>的简单策略。可构造“E1 的 R”的中文形式 query，或“the R of E1”的英文形式 query。显然，预期的搜索结果为 E2。例如，对于三元组<安徽，别名，<a>皖（简称）> ，可以得到“安徽的别名”进行搜索。

基于两个三元组连接的策略。例如，对于三元组<中国， 首都， 北京>和<北京， 面积， 2154.00 万人(2018 年)> ，可以构造“中国的首都的人口”进行搜索。

基于基于模板学习的策略。例如，已有三元组<中国，最高的山，喜马拉雅山>，我们可以从中学习出关系“最高的山”，那么对于所有形如<地名，R，山>的三元组，可以生成查询“地名的最高的山”。

对于一次“搜索”返回的结果，可能出现以下情形：给出直接的答案（图 1）、返回精选摘要（图 2）、只返回相关网页。

图 1

图 2

给出直接答案的 query 根据答案是否正确还可更具体地分为：答案正确、答案错误、答案不准确 3 类。其中出现答案不准确的情形主要是由于答案过时或者存在精度问题。

返回精选摘要的 query 还可以更具体地分为：结果隐含答案，但并非直接给

出答案与结果无答案两种情况。

3、实验设计

3.1 缺陷类型统计与分类

实验使用的知识库数据集为中文通用百科知识图谱（CN-DBpedia），其中包含超过 6500 万条三元组。CN-DBpedia 主要从中文百科类网站（如百度百科、互动百科、中文维基百科等）的纯文本页面中提取信息，经过滤、融合、推断等操作后，最终形成高质量的结构化数据，供用户使用。为提高自动生成 query 的效率，所有三元组数据被分开存储到 14 个生成器中。

首先，基于一个三元组<E1,R,E2>的简单策略，随机生成 query 并根据返回的结果分析缺陷类型。实验随机生成了 4000 次中文 query，所有 query 都通过百度搜索引擎进行查询，并将结果保存。我们将重点关注返回结果中给出直接答案的 query，并将答案错误和答案不准确类型的结果进行了分类。结果中给出直接答案的 query 有 1108 次，具体缺陷分类如下：

答非所问的类型。这一类的缺陷具体表现为返回的答案所回答的问题与 query 不一致，如图 3 所示，query 为“10.30 四川达州被封煤窑中毒事件的发生地”，而返回的答案则回答了“10.30 四川达州被封煤窑中毒事件的时间”。这种类型的缺陷在实验中出现了 4 次。

图 3

query 本身存在语义问题或表义不明。这一类的缺陷由于语义问题被认为不应该或不回返回结果，但是被测搜索引擎事实上却给出了直接答案。例如，“上海市衡山度假村的火车站距离”。这一类型的缺陷在实验中出现了 24 次。

返回奇怪的答案。这一类的缺陷表现为搜索引擎给出的直接结果明显错误且形式奇怪。例如，下图 4 给出了一个返回奇怪答案的例子。这种类型的缺陷在实验中出现了 4 次。

图 4

由于精度问题造成的搜索引擎给出的直接答案不准确。这种类型的缺陷在实验中出现了 10 次。

由于知识图谱过时而导致的错误。这种类型的缺陷在实验中出现了 13 次。

query 不满足“唯一性原则”。这一类缺陷主要是因为为主体本身存在歧义导致的，其问题出在 query 本身。搜索引擎只给出了 query 一种情况下的答案。最明显地表现为人物“重名”问题。例如：“丁薇的生日”、“姚明的身高”等。这种类型的缺陷在实验中共出现了 142 次。

3.2 更改 query 形式

该实验旨在通过在不改变 query 意思的前提下变更 query 形式，比较搜索引擎两次返回的结果，分析 query 形式对语义搜索引擎给出直接结果能力的影响。

实验基于 SwapJD 蜕变关系的思想“一个稳定的搜索引擎应该为相似的查询返回相似的结果”构造 source query 与 follow up query。根据所给三元组<主体，属性，属性值>构造的 source query：“主体的属性”，且预期结果为属性值，现构造 follow up query 形式为：“属性 主体”。

直观上来说由于 source query 的形式更符合汉语的表达方式因此更容易返回结果。 但是在不考虑词序的情况下，根据 tf-idf 算法思想，两者返回的结果应相同或者说结果应存在很大的相似。具体实验结果将在下一章节讨论。

3.3 进一步约束 query 主体

该试验目的在于 1）只根据 source query 得到的结果去分析搜索引擎返回的结果是否与预期的一致 。2）根据 source query 和 follow up query 做一个对比，分析在增通过进一步约束实体来缓解甚至消除实体搜索“不具有唯一性”的影响后，搜索引擎能否仍然给出相同的结果。

本次实验选择了知识图谱中具体的人物类型主体，原因在于人物类型数量庞大且人物类型主题中的属性类型相似，降低了检索同类型实体与大量构造 query 的难度。

实验首先需要从知识库中检索出可能为人名的主体，具体方法为判断主体是否具有“国籍”和“出生日期”两种属性。其国籍还可以作为中国人物实体和外国人物实体分类的依据，出生日期为 query 的预期结果。在检索出可能的人物实体后，同时保存其“职业”属性的值，用于构造 follow up query。在上一步所有检索出的实体中进一步检索出“国籍”为“中国”的实体（属性值中包含“中国”或“中华人民共和国”），最终得到了实体个数为 323840，更近一步，对得到的实体进一步检索其“职业”属性，得到职业属性不为空的实体数目为：146760。

source qu ery 格式为：“主体的生日”， 预期结果即出生日期属性的值。 如果职业属性不为空，则可以构造 follow up query：“职业”+“主体的生日”。此外知识图谱对于人物类型实体重名情况的处理方式为：在人名后添加该人物实体的身份解释。因此若职业属性为空值，还可根据人名后的解释来构造 follow up query。具体实验结果将在下一章节讨论。

4、实验结果

4.1 更改 query 形式后的实验结果

本次实验从知识库中选择任意实体随机地构造了 2032 个 query，并最终成功在百度搜索引擎上进行了 1915 测试。小部分 query 测试失败的原因在于网络不稳定或频繁搜索导致的访问站点超时。对于 source query 中，有 563 次搜索引擎给出了直接结果，有 20 次以精选摘要的形式给出结果。而对于 follow up query

而言，有 42 次搜索引擎给出了直接结果，27 次返回了精选摘要形式的结果。使用“属性 主体”这种形式的 follow up query 进行测试时，被测搜索引擎给出直接答案作为返回结果的次数明显要少于 source query 形式，而返回精选摘要形式的结果次数则稍多。

可以发现搜索引擎对关键词顺序是敏感的，另外符合汉语习惯的 query 形式更容易被认为是实体搜索因此更容易返回精确结果，而对于不符合汉语习惯的 query，搜索引擎只根据关键词排序算法返回了相关的网页。这与直接的推测大致相同。

更进一步，输出了 follow up query 中返回精确结果的 query，无一例外，这些 query 的主体都是与具体人物类型实体，例如“孙成哲的出生日期”、“赵锋佩的职业”。

4.2 进一步约束 query 主体实验结果

实验最终进行了 2134 次成功的 source query 测试，其中有 1174 次搜索引擎给出了直接的答案，7 次以精选摘要的形式返回结果。由于出生日期格式的不一致问题（如：“1996 年 1 月 8 日”、“1996.1.8”、“1996-01-08”），将出生日期结果进行统一化后与知识图谱中保存的结果比较发现结果答案不一致的次数为 249 次，这说明搜索引擎对于 query 的理解与用户的期望存在偏差。

通过 source query 的主体最终生成了 1039 次增加“职业”属性的 follow up query，用 follow up query 对百度搜索引擎进行测试后，只有 1 次测试给出了直接答案，没有出现以精选摘要的形式返回结果的情形。因此，在添加职业属性进一步约束 query 主体后，并没有提升用户体验，反而严重降低了搜索引擎返回结果的能力。

5.总结

本次实验主要基于知识图谱与蜕变关系对搜索引擎进行了相关测试。旨在通过测试发现搜索引擎中存在的一些缺陷，同时分析了变更 query 形式与进一步约束 query 的主体对搜索引擎返回结果的影响。

本次实验仅对百度搜索引擎作为被测搜索引擎，接下来将对 Google 与 Bing 做进一步测试，以分析不同搜索引擎间的能力差异。