基于句子匹配的文章自写度测评系统

化柏林

（中国科学技术信息研究所 北京 100038）

【摘要】 针对人工进行不同文章中相同内容的判断存在着较大困难的局面，提出一个基于句子匹配的文章自写度测评系统。设计基于句子匹配的文章自写度测评系统的系统结构，论述句子分析器、句子匹配器与文章自写度评价器 3 个关键模块的详细流程，并设计相应算法。选取小规模数据进行实验，实验结果表明，基于句子匹配的文章自写度测评系统从技术上完全可行。最后分析基于句子匹配的文章自写度测评系统的难点及问题。

【关键词】 参考文献 内容分析 句子匹配 辅助审稿 句子相似度 【分类号】 TP391 G31

Article Novelty Evaluation System Based on Sentence Matching

Hua Bolin

(*Institute of Scientific and Technical Information of China*, *Beijing 100038*, *China*)

【**Abstract**】 This paper constructs a Article Novelty Evaluation System based on Sentence Matching (ANES – SM), aiming to overcome the difficulty of recognizing same contents between an article and other articles manually. Architecture of ANES – SM is built, and definite flow of key module is analyzed and algorithm is designed, inclading sentence analyzer, sentence matcher and article novelty evaluator. Experiment shows that it is feasible.

【**Keywords**】 Reference Content analysis Sentence matching Aided – review Sentence similarity

1 引 言

句子匹配分析来源于机器翻译，在自动摘要与自动问答等领域得到了迅速发展，但在编辑校对、辅助审稿以及学术传承等领域尚未得到足够的重视。相似句子判定的研究主要应用在基于实例的机器翻译[1-6]、自动文摘[7,8]、信息抽取与自动问答系统[9-11]、编辑校对与辅助审稿[11-14]等。相似句子判定主要考虑词形相似度、词序相似度、句型结构相似度等指标，以及在此基础上构建的向量空间法、依存结构法、编辑距离法。无论哪种方法、应用于哪个领域，分词都是必不可少的第一步，有的系统还需要进行词性标注甚至句法归约；在涉及语义计算的系统中，大都使用语义词典或者知识库。

句子是组成文章的重要单位，也是表明作者行文观点的最小单位。因此把文章切分成句子进行内容分析颇有意义。编辑部收到一篇稿子，利用句子匹配分析可以得到文章的自写度（自写不一定为创新，但相同可能为抄袭或引用）。对每一个句子都有匹配度，编辑会一目了然地看清有哪些句子是抄的，哪些句子是参考别人的，哪些句子是自己写的。现在的编辑部大都依靠人工去判断，人工所能阅读的范围太有限了。如果有这样一个辅助系统能够帮助编辑部人员进行句子级的查找，可以进一步提高编辑部的审稿速度与质量。笔者在参考前人成果的基础上，开发了一个基于句子匹配的文章自写度评测系统 ANES – SM，算法采用 Java 在 Eclipse 下开发，用户界面采用 JSP 实现，数据库采用 Oracle8.1.6。所有功能模块皆为自行开发，包括中文分词器。

收稿日期：2007 – 07 – 12
收修改稿日期：2007 – 09 – 03

2 文章自写度测评系统的总体设计

基于句子匹配的文章自写度判定系统由句子识别器、句子分析器、句子存储器、句子搜索器、句子匹配器和文章自写度评价器 6 个功能模块组成。句子识别器从文章中按照切分单元把文章内的所有句子识别出来，过滤掉不可能含作者观点的小句子，把识别出来的有效句子送交句子存储器。句子存储器负责为句子分配编码，建立句子与文献的顺排档与倒排档索引。句子分析器对识别出来的句子进行各个层面的分析，包括词法分析、语法分析，甚至语义分析，建立句子的各级结构。这些结构包

括格式化句子、句子关键词序结构、句子表层结构与句子深层结构。句子搜索器搜索要匹配的句子,包括搜索范围的确定与搜索顺序的选择。句子匹配器计算目标句子与源句子之间的相似度。相似度可以从符号形式、语法结构、语义表达 3 个层面进行计算。文章自写度评价器计算所有句子的新颖度,并根据句子的权重计算整篇文章的自写度。基于句子匹配的文章自写度判定系统模块结构如图 1 所示:

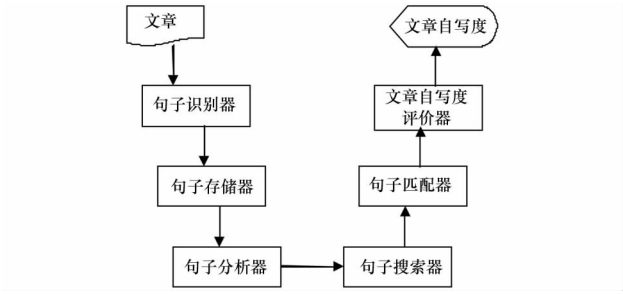

图 1 基于句子匹配的文章自写度判定系统模块结构图

句子识别器、句子存储器与自写度评价器涉及句子与文章之间的操作,句子识别器从一篇文章中识别出 M 个句子,句子存储器建立句子与文章的对应关系,文章自写度评价器根据句子权重与句子新颖度判定文章的自写度。句子分析器、句子搜索器和句子匹配器都是句子与句子之间的操作,句子分析器对句子逐个进行处理,句子搜索器与句子匹配器都是一句对多句的 1:M 型操作。句子分析器、句子匹配器与文章自写度评价是本系统的重点模块。

3 句子分析器

句子分析器对句子进行各个层面的分析,以备后续的句子匹配所用。利用关键词表对句子进行切分,滤掉标点符号和停用词后,得到关键词句子库(每个句子由关键词序列组成),然后进行词性标记、词汇义项标记,接下来进行句法分析、语义分析,得到句子深层结构。句子分析模块处理流程如图 2 所示,其中虚线内容表示可选。

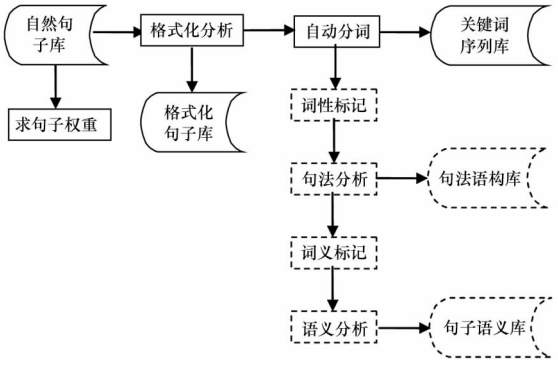

图 2 句子分析器处理流程图

句子的语形分析即在保留句子的原始表达的基础上,对部分内容作表达层面的处理,也就是对句子进行格式化处理。这种格式化处理负责过滤掉句内不重要的信息,对有多种表达方式的部分进行归一化处理。主要包括括号内容的过滤、“的”类字的过滤,英文大小写与标点全半角符号的归一,英文单词、缩略语与中文词汇的归一,同义词的归一处理。

句子的语法分析也称句子的浅层结构分析。句子的浅层结构分析一般指根据词性标记的结果进行句法归约,把词的组合标记成更高层面的短语。本系统进行分词与词性标记时发现,用关键词作词表,使用最大向量分词策略把名词型关键词短语直接标记出来,不进行嵌套分词,然后由这种关键词序列构成句子的顺排档,亦称句子的主干结构,对匹配分析非常有利。短句子所分出的关键词序列太短,会造成过度的相关性判定。例如,“数据挖掘也就是知识发现”与“数据挖掘不同于知识发现”,这两个句子的顺排档一致,但语义却完全相反,如图 3 所示:

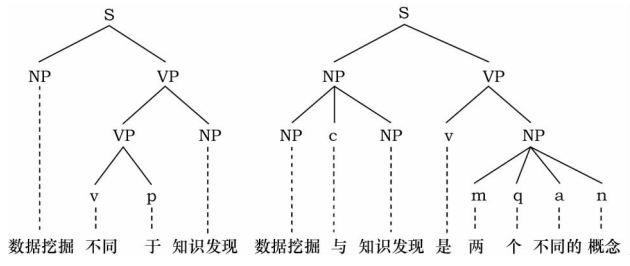

图 3 句子浅层结构示意图

句子的语义分析是句子的深层结构分析。在乔姆斯基的标准理论中,句法部分为每个句子规定深层结构和表层结构;深层结构深入到语义部分,通过语义分析得到句子的语义表达[15]。如“数据挖掘不同于知识发现”和“数据挖掘与知识发现是两个不同的概念”这两个句子的表层结构不同,前者是 S(NP<数据挖掘>,VP(VP(v<不同>,p<于>),NP<知识发现>))①,后者是 S(NP(NP<数据挖掘>,c<与>,NP<知识发现>),VP(v<是>),NP(m<两>,q<个>,a<不同的>,n<概念>)),成分结构如图 3 所示。两个句子的深层结构是相同的,都为 different<数据挖掘,知识发现>,句子的深层结构通常以逻辑形式来表达。这样,有些作者变换句型而不变句子语义的话,仍然能够分析出来,如主动句与被动句、把字句与被字句、带介词的双宾语句与不带介词的双宾语句等。不同的表述不能构成创新,但是不同的实

① 分词时使用关键词词库,遇到名词短语型关键词(如数据挖掘、知识发现等)不进行细分,以保证词的完整性、分词速度与句法规则的鲁棒性。

验数据,不同的结论和不同的方法就是一定程度的创新。

4 句子匹配器

句子匹配器主要负责计算目标句子与源句子之间的相似度。相似度可以是二值判断,也可以是多值判断。二值判断即要么匹配、要么不匹配两种结果。而多值判断是可以给出任意两个句子的相似度。本系统中把相似度的计算分为以下几种情况:

(1)全字符比较:对文章中抽取出的每一个句子,直接进行全字符比较。全字符匹配的结果直接标以 100%。这种方法速度较快,但是准确率较低,对原装复制型抄袭较为有效,对经过编辑改动的情况则无能为力。

(2)格式化句子匹配:格式化句子即对原始句子表达进行处理,过滤掉无用词语及对部分词语进行归一化处理的句子。这种匹配一般是文字层面,即语形级的匹配,如果匹配,结果标记为 99%;如果不匹配,则视为第三种情况,进入模糊匹配处理过程。

如果句子全字符与格式化句子皆不匹配,则需要进行句子主干匹配。句子主干匹配是根据句子顺排档搜索到匹配的句子,然后对句子的原始结构进行相似度的计算。计算时从词的层面进行匹配,包括词形与词序,判断其相似性。句子匹配器的流程如图 4 所示:

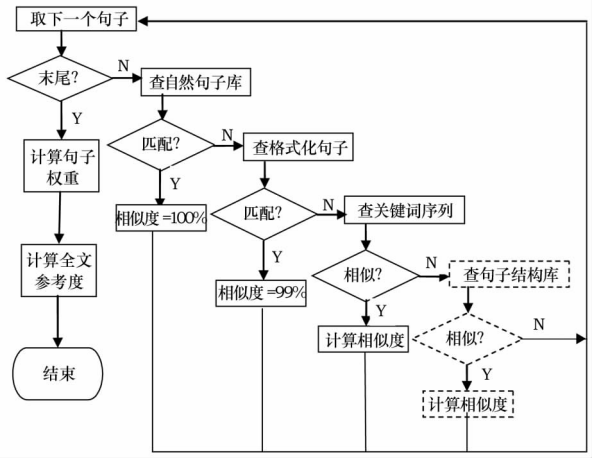

图 4 句子匹配器流程图

对句子 S_i 来讲,每遇到一个顺排档相同的句子,都会计算出一个相似度,取相似度最高的那个句子为最终的结果。计算结果主要考虑匹配词的个数、每个词的长度以及两个句子的长度。对任意两个顺排档相同的句子,其相似度等于所有匹配词的长度和除以两个句子长度和的一半,计算公式如下:

$$Sim(S_i, S_j) = \frac{\sum_{k=1}^{n} L_k}{(Len(S_i) + Len(S_j))/2}$$

L_k 为两个句子中第 k 个相同片段(可能是字、词、短语或者它们之间的组合)的长度,n 为相同片段的个数,即任意两个句子之间的相似度等于相同片段的长度和除以两个句子的平均长度。对于长句与短句的包含关系,特别是某一句包含在另一大句中的情况,采用顺排档包含的匹配方式,而不是采用相等进行匹配。

5 文章自写度评价器

判断一篇文章的自写程度有以下 3 个指标:

(1)自写率:自写率是用全文句子总数 M 减去雷同句子数 N,然后再除以全文句子总数 M 得到文章的自写率$\frac{M-N}{M}$。

(2)比较句均参考度:所有句子的参考度累加除以总句子数,同自写率一样,这种算法假定所有句子对文章贡献率一样,而不考虑句子之间的不同。

(3)加权自写率:不同位置的句子有着不同重要性,不同长度的句子重要性也不一样,就是要充分考虑句子对文章的贡献度。主要考虑两个方面,一方面是句子的位置权重;一方面是句子的本身信息,包括句子的长度和复杂度。句子越长,对文章的贡献率就越大;句子越短,对文章贡献率就越小。句子复杂度越高,对文章的贡献率就越大;句子复杂度越低,对文章的贡献率就越小。文章自写度的计算公式如下:

$$S = \frac{\sum_{i=1}^{n} (N_i \times P_i + \frac{L_i}{L})}{n}$$

公式中,n 为句子的总数,N_i 为句子的新颖度,P_i 为句子的位置权重,L_i 为句子的长度,L 为平均句长,文章自写度 S 就等于所有句子的新颖度乘以句子的权重,累加求和以后除以句子总数。

相似度越高,引用或抄袭别人的可能性就越大;相似度越小,说明该句子作者自写的成分就越高,创新性就可能也越强。如果句末没有作标引且相似度大于 90% 的句子可以判定为抄袭句。相似度为 100% 的句子是显性抄袭;相似度为 99% 的句子是隐性抄袭(经过同义词替换处理过的句子);介于 90% 到 98% 的句子为潜在抄袭。如果在这些句子的末尾作了参考文献的标注,就变成了显性引用、隐性引用(作了标注就没有潜在引用一说了)。除了这些指标外,还可以用抄袭量、引用量与匹配量 3 个指标来计算作者的信用指数。

6 实验结果与分析

实验分 3 个阶段进行,3 个阶段的自动化程度越来越高,数据规模越来越大,分析层面越来越深,对系统验证的可信度越来越高。第一阶段人工选择特定文献进行实验;第二阶段对聚过类的文献进行分析,文献的数据量上升到几百篇至几万篇,数据量较大。第三阶段对大规模文献进行分析。

目前的实验正处于第一阶段。这一阶段的实验有 3

个针对性的实验,实验一分析抄袭现象;实验二分析一稿多发现象;实验三分析合成稿现象(由自己的几篇相关文章组合成一篇新文章)。实验一的实验已经完成,实验二与实验三的文献已选好。

实验一以 M7302117610050302 为处理内容,经过计算发现,M7302117610050302 的内容绝大部分来自 J100335132004S115、J1003031x200009033 等 4 篇文献。全文没有一处实标注,而且这几篇文章也未列在参考文献目录里。文章 M7302117610050302 含有 111 个句子(滤掉小句以后),其中完全匹配的句子有 37 个,基本匹配的句子(经过中英文及大小写处理过的句子)有 17 个,相似度大于90% 的累计有 72 个,约占句子总数的 2/3,各个句子的匹配结果见表 1。系统运行结果如图 5 所示。

表 1 第一阶段实验一的句子匹配结果

匹配度(%)	句子数	累积百分比(%)
100	37	33.33
90 – 99	35	64.86
80 – 89	10	73.87
70 – 79	8	81.08
60 – 69	2	82.88
50 – 59	10	91.89
40 – 49	3	94.59
30 – 39	5	99.10
20 – 29	1	100.00
10 – 19	0	100.00
<10	0	100.00

图 5 系统运行结果图

人工校对发现,M7302117610050302 绝大部分句子都来源于其他文献,但作者进行了很多编辑加工,特别是中英文词语的替换、英文大小写及缩略语的替换以及同义词的替换、增删小字词以及大小句的拆分与组合。

7 结 语

基于句子匹配的文章自写度测评系统已进行了初步实验,但离大规模应用还有一定的距离,主要体现在以下 3 个方面:异构数据的获取;历史数据的回溯建库和跨语言之间的判定。一篇文章的参考文献有很多种形式,其中常见的有期刊论文、学位论文、会议论文、图书、报告等,当然还有网页、报纸等形式。目前,没有一家数据库商能提供这十大文献源,有的能提供期刊论文和学位论文,如清华同方、万方数据等;有的能提供电子图书,如超星电子图书馆,Apabi 图书等;有的能提供网页,如 Google、百度等,目前还很难把它们的数据集成到一起。第二个难点是历史数据的获取问题。如引用的文献比较早,特别是对原始文献的引用要求越来越高,有些文献很难获取电子版,句子匹配也就不现实了。但是,中国知网的很多期刊已经回溯至创刊号,这一问题也能得到一定的解决,全文扫描与识别工作也会进一步加强。第三个难点是跨语言之间的判定问题。文章引用其它语种的文献比较多,如中文文章引用英文文献,硕博士学位论文的英文参考文献数量超过参考文献总数的 1/3,目前的技术还很难分析出中文句子与英文句子的匹配,但是可以通过建设双语语料库来逐步解决这个问题。

基于句子匹配的文章自写度测试系统 ANES – SM 不能解决结构性抄袭的问题。结构性抄袭指不同领域或不同主题的论文之间存在着结构与语言表达的抄袭,在研究方法、结论、论文结构、语言表达等方面完全一致的基础上,换成自己的数据与主题词,这类抄袭以主题型计量分析类文章居多。系统能给出每个句子的相似度,但相似度达到多少就是抄袭,这个阈值的确定目前还没有解决。

总之,语言现象太复杂,想设计一个通用的程序能解决所有的语言问题是不现实的,也是不可能的。随着数据量的不断加大,分析层面的不断加深,也会遇到越来越多的问题,这也需要程序的鲁棒性越来越强,算法的速度越来越快。只有不断积累资源(包括好的词典与规则),不断改进算法,逐步提高系统的健壮性与通用性,才能更好地满足应用需求。

参考文献:

[1] 吕学强,任飞亮,黄志丹,等. 句子相似模型和最相似句子查找算法[J]. 东北大学学报(自然科学版),2003,24(6):531 – 534.

[2] 姚建民,周明,赵铁军,等. 基于句子相似度的机器翻译评价方法及其有效性分析[J]. 计算机研究与发展,2004,41(4):1258 – 1265.

[3] 王荣波,池哲儒. 基于词类串的汉语句子结构相似度计算方法[J]. 中文信息学报,2005,19(1):21 – 29.

[4] 王荣波,池哲儒,常宝宝,等. 基于词串粒度及权值的汉语句子相似度衡量[J]. 计算机工程,2005,31(13):142 - 144.
[5] 黄河燕,陈肇雄,张孝飞,等. 大规模句子相似度计算方法[J]. 中文信息学报,2006,20(S1):47 - 52.
[6] 林贤明,李堂秋,陈毅东. 句子相似度的动态规划求解及改进[J]. 计算机工程与应用,2004,40(35):64 - 65,93.
[7] 李彬,刘挺,秦兵,等. 基于语义依存的汉语句子相似度计算[J]. 计算机应用研究, 2003,20(12):15 - 17.
[8] 张琦,黄萱菁,吴立德. 一种新的句子相似度度量及其在文本自动摘要中的应用[J]. 中文信息学报,2005,19(2):93 - 99.
[9] 秦兵,刘挺,王洋,等. 基于常问问题集的中文问答系统研究[J]. 哈尔滨工业大学学报,2003,35(10):1179 - 1182.
[10] 廉站俊,吕学强,张玉杰,等. 基于句子相似度计算的信息抽取[J]. 现代图书情报技术,2007(6):38 - 41.
[11] 金博,史彦军,滕弘飞. 基于语义理解的文本相似度算法[J]. 大连理工大学学报,2005,45(2):291 - 297.
[12] 车万翔,刘挺,秦兵,等. 基于改进编辑距离的中文相似句子检索[J]. 高技术通讯,2004,14(7):15 - 19.
[13] 郑逢斌,陈志国,姜保庆,等. 语义校对系统中的句子语义骨架模糊匹配算法[J]. 电子学报,2003,31(8):1138 - 1140.
[14] 李卫,王极,李蕾,等. 全信息知识制导的科技期刊初审辅助系统[J]. 北京邮电大学学报,2006,29(S1):127 - 132.
[15] 刘颖. 计算语言学[M]. 北京:清华大学出版社,2002:45 - 46.

(作者 E - mail:huabolin@ istic. ac. cn)

动 态

全球科学门户网站开启

美国能源部、英国图书馆以及其他 8 个参与国在华盛顿于 2007 年 6 月 22 日共同开启了从全球 15 个国家入口接入的全球在线科学信息门户。此信息门户的网址为 http://www. WorldWideScience. org,它可以为普通市民、研究人员以及任何对科学感兴趣的人提供科学信息的搜索入口,以便他们能够轻松访问那些使用普通搜索技术不能访问的网站,例如 Google、雅虎和许多其他商业搜索引擎所使用的搜索技术。

美国能源部科学部副部长 Raymond L. Orbach 说:“在互联网上大量的信息资源中,许多信息都无法通过普通的搜索引擎获知和被检索到,而通过全球科学搜索这些信息都能被搜索和获得,并且这样国际性的合作将通过一种快捷和便利的方式创建一个巨大的知识库,这将使得我们现有的知识获得更大程度的增值”。

WorldWideScience. org 门户网站使用联合搜索的创新技术,使用者一次查询时就可以同时搜索国际国内的科学入口网站,同时也能搜索到一般商业搜索引擎无法搜索到的资料。

WorldWideScience. org 门户入口还为澳大利亚、巴西、加拿大、丹麦、法国、德国、日本和荷兰提供了英语信息搜索权限。创建 WorldWideScience. org 的目的在于使其成为世界级的网络工具,让任何地方的任何一位科学家或任何人都能轻松地获得任一种语言或任一国家的搜索结果。

WorldWideScience. org 是美能源部科学办公室的一个项目,由美科技情报办公室开发的,现在该机构也负责它的维护工作。

(编译自:Global Science Gateway Now Open. [2007 - 07 - 18]. http://www. doe. gov/news/5153. htm.)

(本刊讯)