

doi:10.11835/j.issn.1000-582X.2019.09.013

# 面向自动问答系统的问句相似度计算研究

曹建文<sup>1</sup>, 万福成<sup>2</sup>

(1. 兰州工业学院 软件工程学院, 兰州 730050; 2. 西北民族大学 中国民族信息技术研究院, 兰州 730030)

**摘要:** 结构化自动问答系统采用传统方法缺少对词汇、词序和结构的划分, 导致语句相似度较低, 为了解决该问题, 提出了基于 Web 语义的混合问句相似度计算方法。根据结构化自动问答系统结构, 设计系统语句分析模型, 通过正向匹配方法, 对模型专业词库中的用户输入自然语句进行分词处理, 并对字符串之间的关系展开分析。采用非恒定相似度系数来描述 2 个字符串的相似情况, 并由此分析词形、词序和结构相似度, 完成不同语句相似度的计算。通过实验对比可知, 文章提出的基于 Web 语义的混合问句相似度计算方法最高计算精准度可达到 96%, 可提升自动问答系统的整体性能。

**关键词:** 结构化; 自动问答; 相似度

**中图分类号:** TP311.51

**文献标志码:** A

**文章编号:** 1000-582X(2019)09-114-09

## Question similarity computing method for automatic question answering system

CAO Jianwen<sup>1</sup>, WAN Fucheng<sup>2</sup>

(1. College of Software Engineering, Lanzhou Institute of Technology, Lanzhou 730050, P. R. China;

2. National Language Information Technology, Norhewest Minzu Univesety, Lanzhou 730030, P. R. China)

**Abstract:** Structured automatic question answering system lacks the division of vocabulary, word order and structure, which leads to low sentence similarity. In order to solve this problem, a method of computing sentence similarity based on Web semantics is proposed. According to the structure of the structured automatic question answering system, the system statement analysis model is designed. Through the forward matching method, the user input natural statements in the model professional lexicon are segmented and the string relationship is analyzed. The unsteady similarity coefficient is used to describe the similarity of two strings, and the similarity of morphology, word order and structure is analyzed to determine the similarity of different sentences. The experimental results show that with this method, the minimum computational accuracy is 42%, and the maximum computational accuracy can reach 96%. The overall computational accuracy has been greatly improved compared with that based on neural network and ant colony, thereby improving the overall performance of the automatic question answering system.

**Keywords:** structuration; automatic question answering; similarity

**收稿日期:** 2019-03-12

**基金项目:** 国家自然科学基金青年资助项目(61602387)。

Supported by National Natural Science Foundation for Young Scholars of China (61602387).

**作者简介:** 曹建文(1972—), 男, 副教授, 主要从事计算机技术研究, (E-mail) wanfucheng@126.com。

**通讯作者:** 万福成, 男, 副教授, 博士, 主要从事自然语言处理、自动问答技术和信息抽取研究, (E-mail) 306261663@qq.com。

自动问答系统是目前自然语言处理系统的核心任务之一,近年来受到越来越多的学者关注,它能使用户利用自然语句提问,为每一个用户获取简洁、准确的答案。传统系统是采用关键字搜索方法,使用户能够快速找出答案,满足用户检索需求<sup>[1]</sup>。结构化自动问答系统将用户提出的问题,转换为计算机可以理解的方式,可通过查找相似语句来完成。现有问答系统在计算语句相似度方面具有如下问题:组合词汇语义相似度不能准确计算;缺少对非专业词汇相似度相关计算;在分词方面借助词典进行简单区分,不能准确划分专业词汇以及组合词汇<sup>[2]</sup>。在自然语言处理领域中,句子相似度计算是一项核心研究课题,其研究状况直接决定相关领域的研究进展,对各个领域都非常重要。

目前句子相似度计算方法主要有 3 个方面。第一个方面是以词语为粒度的计算方法,例如,基于词语重叠的方法<sup>[3-4]</sup>,基于词语语义的方法<sup>[5-7]</sup>,向量空间方法<sup>[8]</sup>等。词语重叠的方法主要是通过统计 2 个句子中共有词语的数量进行分析,词语语义方法主要通过构建词语的相似度矩阵来计算词语的语义相似度,向量空间方法即 SVM(support vector machine)方法,通过 SVM 来计算问句语义相似度。第二个方面是以句法为特征的计算方法,例如,Li 等<sup>[9]</sup>将汉语的依存关系句法信息融入到问句分析中,Oliva 等<sup>[10]</sup>将句法角色信息融入到问句分析模型,在句法层面,Chang 等<sup>[11]</sup>将长句子切分成短句,再进行问句分析。第三个方面是混合的方法,例如 Saric 等<sup>[12]</sup>采用了词语重叠、句法相融合的方法,熊大平等<sup>[13]</sup>提出一种基于 LDA 的问句相似度计算方法,田卫东等<sup>[14]</sup>引入问句类别特征记性问句相似度计算方法。这些方法虽然取得了一定效果,但仍然受到问答系统性能的影响,不能保持较高计算精度,针对该问题,提出了基于语义 Web 的混合问句相似度计算方法,并使用该方法在问答系统中进行了应用,验证了该方法的合理性。

### 1 结构化自动问答系统结构

结构化自动问答系统可用来分析用户提出的问题,通过结构化方式来找寻答案,如果可以直接在数据库中找到相似问句,则直接从该问句的所有答案中提取出想要的信息,系统结构如图 1 所示。

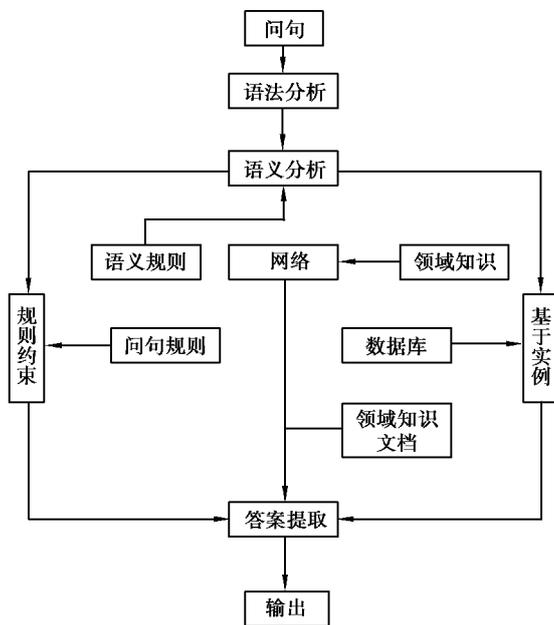


图 1 结构化自动问答系统结构

Fig. 1 Structure of structured automatic question answering system

由图 1 可知:当用户提交疑问时,需先对问题进行分词处理。分词处理主要有基于规则推理和基于实际推理 2 种处理机制。答案提取模块是通过对语义描述进行综合处理,由此能够获取很多答案<sup>[15]</sup>。

## 2 基于语义 Web 的混合问句相似度计算

根据结构化自动问答系统结构,设计自动问答系统模型,通过该模型,计算基于语义 Web 的混合问句相似度<sup>[16]</sup>。

将自动问答系统模型分为用户显示层、中间分析层和数据获取层 3 层结构,其中用户显示层是将用户提出的问题显示在系统上;中间分析层负责分词、处理和计算语义相似度;数据获取层主要从专业词库、同义词库和领域词库中获取数据。自动问答系统语句分析模型如图 2 所示。

根据专业词库对用户输入的自然语句,通过正向最大匹配方法进行分词,并借助分词工具将未登记词添加到数据库之中<sup>[17]</sup>。

### 2.1 分词

在结构化自动问答系统之中,问题是针对某一特定领域设计的,比如“二叉树”既是数据结构领域专业词汇,也是复合型词语,而“树”虽然是专业性词语,但不属于复合词。结合图 1 所示的系统结构,采用基于语义 Web 词库,搜集所有专业词语字符串,并采用正向最大匹配算法,标注特征词词性,获取分词权值<sup>[18]</sup>。

由于搜索的词库是不全面的,导致部分语句分词并不理想,针对该问题,一旦出现未登记词语时,需借助现有分词工具对这些词语进行分词处理<sup>[19]</sup>。分词后的权值  $W$  是根据词汇特征来计算的,具体可分为:

$$W = \begin{cases} 0, & \text{停用词汇,} \\ 0.2, & \text{常用词汇,} \\ 0.5, & \text{形容词和副词,} \\ 0.8, & \text{名词和动词,} \\ 1, & \text{疑问词.} \end{cases} \quad (1)$$

由式(1)可知,停用词汇是不影响整个句子相似度的,因此,为了提高系统效率,在分词后应删除停用词汇<sup>[20]</sup>。

### 2.2 字符串关系分析

经过分词处理后,对字符串关系进行分析。计算机中的语句是由不同含义单词组成的,可将其视为一个字符串。综合分析区间标度变量、序数型变量和比例标度变量,假设所有单元词语都属于变量属性集合  $Q$ ,那么字符串  $a$  和字符串  $b$  中所有的单词也都包含在这个变量属性集合。设  $r$  为 2 个字符串中都存在的单词数, $e$  为  $a$  字符串中特殊存在字符, $q$  为  $b$  字符串中特殊存在字符, $z$  为 2 个字符串停用单词总数。 $r, e, q, z$  为字符串中的 4 个状态向量,其间关系如图 3 所示。

结合图 3,采用非恒定相似度系数来描述 2 个字符串相似度,为

$$X = \frac{r}{r + e + q}. \quad (2)$$

问句相似度除了与关键词有关,还与句子长度、关键词顺序有关<sup>[21]</sup>。

### 2.3 语句相似度计算

#### 2.3.1 词形相似度

2 个句子中的词语在形态上相似程度可衡量所有相同词的个数,此时需剔除已经停用的词汇。设  $Y_1, Y_2$  为 2 个句子,那么这 2 个句子相似度为

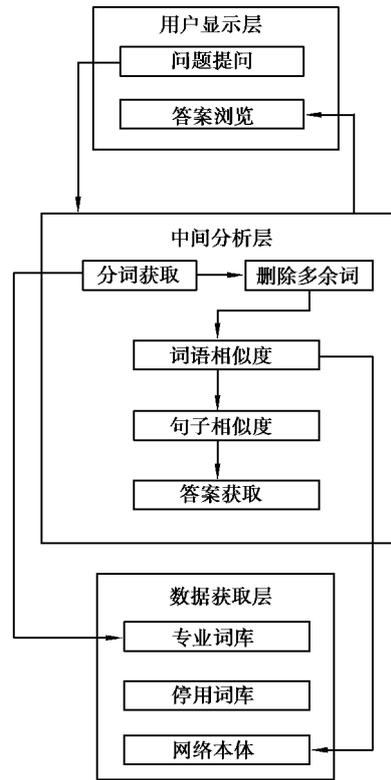


图 2 自动问答系统语句分析模型  
Fig. 2 Sentence Analysis Model of automatic question answering system

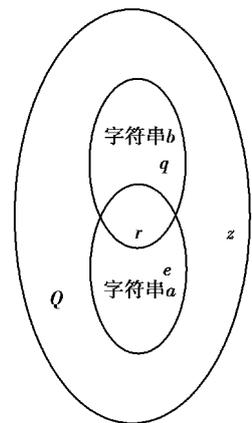


图 3  $r, e, q, z$  为字符串关系  
Fig. 3  $r, e, q, z$  is a string relationship

$$S_{siml}(Y_1, Y_2) = 2 * \frac{S_{sameword}(Y_1, Y_2)}{L_{Len}(Y_1) + L_{Len}(Y_2)}, \quad (3)$$

式中: $S_{sameword}(Y_1, Y_2)$ 表示句子 $Y_1, Y_2$ 中所包含相同词汇数量; $L_{Len}(Y)$ 表示句子 $Y$ 的数量。通过对词形相似度计算,可改善以往方法带来计算机语句相似度词形相似而引起自动回答结果错误的问题。

### 2.3.2 词序相似度

从关键词顺序上标注语句相似度,可直观反映出 2 个语句所包含相同词在位置关系上的相似程度,通过这 2 个句子衡量相同词或同义词相邻顺序逆向个数<sup>[22]</sup>。计算方式为

$$S_{ordsim}(Y_1, Y_2) = 1 - \frac{R_{Rev}(Y_1, Y_2)}{R_{MaxRev}(Y_1, Y_2)}, \quad (4)$$

式中: $R_{MaxRev}(Y_1, Y_2)$ 表示 $Y_1, Y_2$ 相同关键词自然序列中的最大逆序列数; $R_{Rev}(Y_1, Y_2)$ 表示 $Y_1$ 关键词在 $Y_2$ 中所构成自然序列中的逆序列。

衡量 2 个句子中所包含的相同词或同义词相邻顺序逆向关系,需设 $S_{onceWord}(Y_1, Y_2)$ 为相同词集合,重复出现的词仅需计算 1 次,按照关键词出现的先后顺序构成自然数顺序序列,分量按照对应词在第 2 个句子中的次序生成向量。通过对词序相似度计算,可降低以往方法计算量,提高自动回答结果精准度。

### 2.3.3 结构相似度

基于结构化问句实例分析模型如图 4 所示。

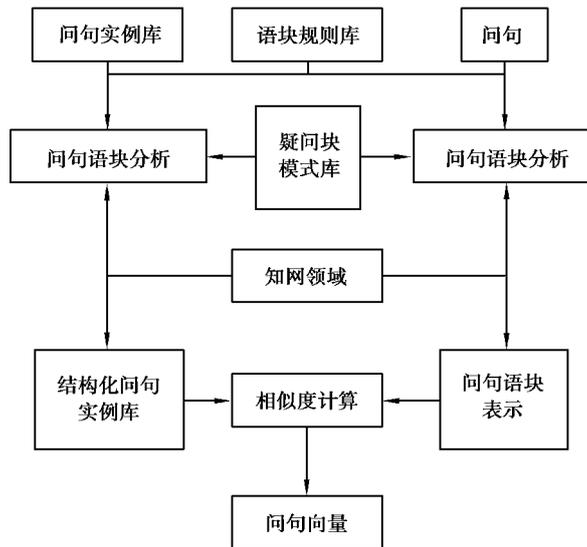


图 4 问句实例分析模型

Fig. 4 Question case analysis model

将问句按照语块组成规则形成有秩序的语句,从问句实例库语句中选取一组可以用来匹配的语块组,通过匹配信息自动分析当前输入问句。通过图 4 所示的问句分析模型,计算语句结构相似度。对结构相似度计算,可快速匹配输入问句,减少自动回答时出现错误结果的机率。

将语句结构按照相似度进行划分,可分成 4 个部分,分别是词形相似度、句长相似度、距离相似度和词序相似度。其中词形相似度表示不同句子相同词语相似的个数;句长相似度表示 2 个问句的长度;距离相似度表示使用不同句子相同词语的距离来标注句子相似度;词序相似度表示相同属性词语的顺序。

问句相似度计算公式为

$$S_{Sim3}(Y_1, Y_2) = \alpha_1 S_{WordSim}(Y_1, Y_2) + \alpha_2 L_{LenSim}(Y_1, Y_2) + \alpha_3 S_{OrdSim}(Y_1, Y_2) + \alpha_4 S_{DisSim}(Y_1, Y_2), \quad (5)$$

式中: $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$ ,  $\alpha$  表示相似度参数。为了保证计算结果更加精准,引进词语权值,对词形相似度和句长相似度进行改进,为

$$S_{\text{WordSim}}(Y_1, Y_2) = \frac{2 \sum_p^i \omega_p}{\sum_n^p \omega_n + \sum_m^o \omega_m}, \quad (6)$$

$$L_{\text{LenSim}}(Y_1, Y_2) = 1 - \frac{\left| \sum_n^p \omega_n - \sum_m^o \omega_m \right|}{\sum_n^p \omega_n + \sum_m^o \omega_m}, \quad (7)$$

式中:  $i$  表示词形数量;  $\omega_p$  表示词形权重;  $\omega_n$  表示句子  $Y_1$  中相似词形权重;  $\omega_m$  表示句子  $Y_2$  中相似词形权重。

实际上,问句中的问答表达方式多种多样,用户输入的问句和计算机数据库中出现完全匹配概率非常低。因此,需确定相似原则,在语义 Web 结构相似情况下,分析词汇和实词相似度。

### 1) 词汇相似度

在研究基于问句实例分析系统背景下,词汇相似度就是 2 个词语在语义上的相似程度。对于 2 个词  $E_1$  和  $E_2$ ,如果  $E_1$  有  $g$  个概念:  $S_{11}, S_{12}, \dots, S_{1n}$ ;  $E_2$  有  $f$  个概念个概念:  $S_{21}, S_{22}, \dots, S_{2m}$ ,  $E_1$  和  $E_2$  相似度是集合各个概念相似度最大值,为

$$S_{\text{Sim}}(E_1, E_2) = \max_{a=1 \dots n, b=1 \dots m} S_{\text{Sim}}(S_{1a}, S_{2b}). \quad (8)$$

在所有词语资源中,最常用的就是知网,其概念最终归结于个别地方具体词,因此,这种具体词相似度计算是概念相似度计算的基础。

对于 2 个个别地方具体词  $A_1$  和  $A_2$ ,需计算其相似度为  $S_{\text{Sim}}(A_1, A_2)$ , 2 个词语之间距离为  $d(A_1, A_2)$ , 由此可通过式(9)转换关系来满足以上条件。

$$S_{\text{Sim}}(A_1, A_2) = \frac{\beta}{d(A_1, A_2) + \beta}, \quad (9)$$

式中:  $A_1$  和  $A_2$  分别表示 2 个个别地方具体词;  $d(A_1, A_2)$  在  $d(A_1, A_2)$  层次体系中路径长度为正整数;  $\beta$  表示可调节参数。

### 2) 实词概念相似度计算

实词概念是由基本个别具体词、关系具体词、符号具体词和其他具体词组成的,将这 4 个部分对应的概念相似度分别记为:  $S_{\text{Sim}1}(R_1, R_2)$ ,  $S_{\text{Sim}2}(R_1, R_2)$ ,  $S_{\text{Sim}3}(R_1, R_2)$ ,  $S_{\text{Sim}4}(R_1, R_2)$ 。在知网中这 4 项具体词描述重要程度依次降低,由此获取的 2 个概念语义表达式整体相似度为

$$S_{\text{Sim}}(R_1, R_2) = \sum_{c=1}^4 K_c \prod_{d=1}^c S_{\text{Sim}}(R_1, R_2), \quad (10)$$

式中:  $K_c$  表示可以调节的参数,并且  $K_1 + K_2 + K_3 + K_4 = 1$ ;  $S_{\text{Sim}1}$  到  $S_{\text{Sim}4}$  总体相似度是呈递减趋势变化的。

根据对词形、词序和结构相似度分析,可实现语句相似度的计算,并避免了以往用户输入的问句和计算机数据库中出现完全匹配概率极低现象发生。

## 3 实 验

为了验证结构化自动问答系统中的问句相似度计算方法研究的合理性,进行了如下实验。

### 3.1 系统执行流程

在目标问句进入结构化自动问答系统之前,需将问句分成词语组合形式,通过建立知识库进行全文搜索,选择相似度最大值,并与设定阈值进行对比分析,具体流程如图 5 所示。

由图 5 可知:从目标句子中抽取分词,进行全文搜索,选择出相似问题,并对其相似度进行计算。如果相似度小于阈值,则说明问题还未被解决,需进行全部信息搜索,并对数据库实时更新;如果相似度大于阈值,则说明问题被解决,此时可给出正确答案。

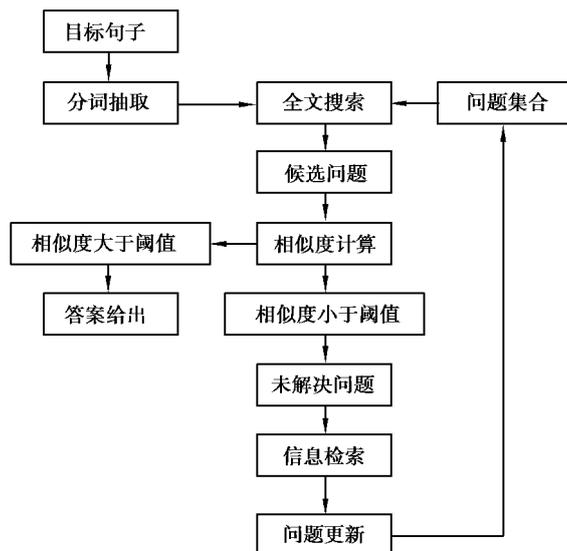


图 5 问答系统流程

Fig. 5 Question and answer system flow

### 3.2 实验问题集建立

在结构化自动问答系统中,将全部问题都存储在 MySQL 数据库之中。首先对用户输入问题进行分词抽取,过滤停用词后,再对关键词进行全文检索,将目标问句中相关语句作为候选问题集,以此构建实验问题集。

### 3.3 答案提取

对于不同问句答案的提取,需根据问句规则在问句中进行分析,引导分析器匹配正确答案。系统通过自动问答机制求解问题,获取表示疑问的问句向量,通常有以下 3 种情况:

- 1) 语句结构+词汇属性+主旨;
- 2) 语句表明原因+词汇扮演角色+主旨;
- 3) 无序关键字。

从所有答案中随机抽取与 1) 情况相关的答案:

- ① 该答案包含知识文档中问点相关词汇;
- ② 虽然该答案缺少具体对应问点词汇,但可表明语句主旨;
- ③ 虽然该答案缺少具体对应问点词汇,但可以找到相似主体概念;
- ④ 该答案缺少具体对应问点词汇,也缺少相似主旨概念。

针对情况①,需直接从文档中提取相关问点。情况②可直接利用属性进行答案提取,具体提取步骤如下所示:设主题概念为  $\varphi$ ,问点属性为  $G$ 。其中  $\gamma \leftarrow S_{upper}(\varphi)$ ,如果  $\gamma$  具有属性  $G$ ,那么可在文档该问点中找到答案;如果没有属性  $G$ ,那么  $\gamma$  就不是根问点,则需重新分析问点属性。记录一次分析结果,并提交给管理员进行维护。针对情况③,通过核心词汇语义相似度进行计算;针对情况④,通过查询语句扩展来处理。

### 3.4 实验结果与分析

根据上述实验内容,对实验结果进行对比分析。保持 2 条语句中词形与词序一致,对比分析基于神经网络计算机语句相似度计算方法、基于蚁群计算机语句相似度计算方法和基于 Web 语义分析方法的语句结构相似度,结果如图 6 所示。

由图 6 可知:基于语义 Web 的计算机语句相似度计算方法与基于神经网络计算机语句相似度计算方法、基于蚁群计算机语句相似度计算方法曲线并不吻合。

保持语句结构和词序一致,分析不同方法词形相似度,结果如图 7 所示。

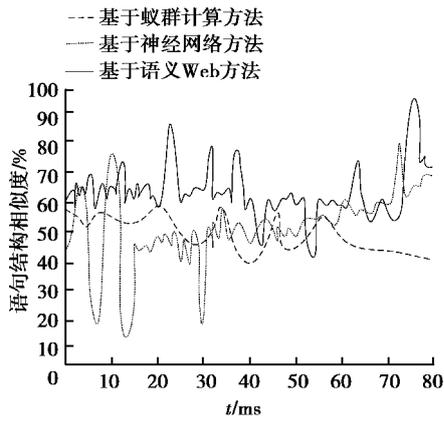


图 6 不同方法语句结构相似度对比分析

Fig. 6 Comparative analysis of sentence structure similarity between different methods

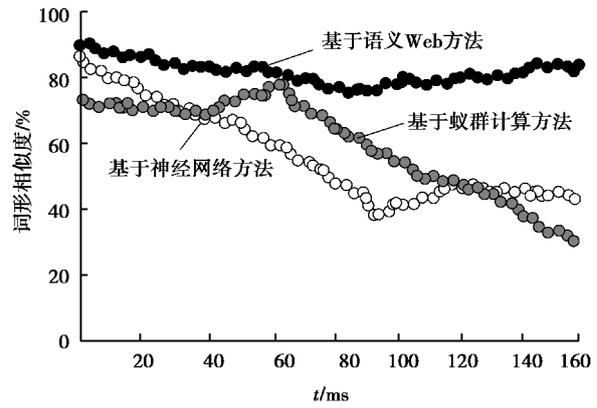


图 7 不同方法词形相似度对比分析

Fig. 7 Contrastive Analysis of word form similarity in different methods

同理,保持语句结构和词形一致,分析不同方法词序相似度,结果如图 8 所示。

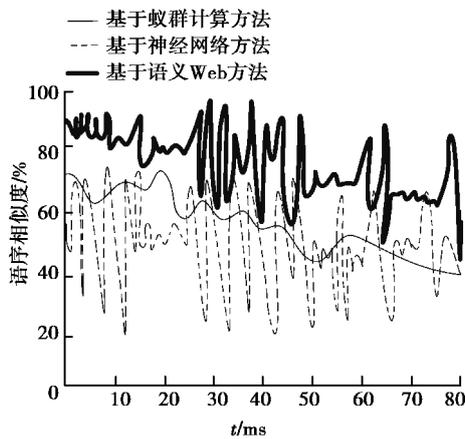


图 8 不同方法词序相似度对比分析

Fig. 8 Contrastive Analysis of word order similarity in different methods

从以上对比实验进行分析,从语句结构、语序、词形 3 个维度进行展示,如表 1 所示。

表 1 实验结论

Table 1 Conclusion of the experiment

%

语句相似度	基于神经网络计算方法		基于蚁群计算方法		基于 Web 语义分析方法	
	最高	最低	最高	最低	最高	最低
语句结构	80	13	58	41	96	42
词序	88	38	71	49	90	78
词形	75	20	89	39	98	55

综上所述:结构化自动问答系统中的问句相似度计算方法研究结果精准度较高,由此可说明该方法是具有合理性的。

## 4 结束语

采用基于 Web 语义的混合问句方法研究结构化自动问答系统中的问句相似度,通过文档整理相关知识,并分析语句中词序、词形和结构相似度,以此为基础,研究语句整体相似度。由实验对比分析可知,基于 Web 语义的混合问句方法是具有合理性的,对于知识表示和管理具有重要意义。

### 参考文献:

- [1] 王丽月,叶东毅.面向游戏客服场景的自动问答系统研究与实现[J].计算机工程与应用,2016,52(17):152-159.  
WANG Liyue, YE Dongyi. Research and implementation of automatic question-answer system in game customer service scenarios[J]. Computer Engineering and Applications, 2016, 52(17):152-159. (in Chinese)
- [2] 李晓,解辉,李立杰.基于 Word2vec 的句子语义相似度计算研究[J].计算机科学,2017,44(9):256-260.  
LI Xiao, XIE Hui, LI Lijie. Research on sentence semantic similarity calculation based on word2vec[J]. Computer Science, 2017, 44(9): 256-260. (in Chinese)
- [3] Metzler D, Bernstein Y, Croft B, et al. Similarity measures for tracking information flow[C]// Acm International Conference on Information & Knowledge Management. ACM, 2005.
- [4] 钟敏娟,万常选,刘爱红,等.基于词共现模型的常问问题集的自动问答系统研究[J].情报学报,2009,28(2):242-247.  
ZHONG Minjuan, WAN Changxuan, LIU Aihong, et al. Question answering system based on frequently asked questions using co-occurrence word model[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(2): 242-247. (in Chinese)
- [5] Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity[C]// The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference. Boston, Massachusetts, USA; DBLP, 2006, 7: 16-20.
- [6] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. ArXiv Preprint ArXiv, 2013, 1301.
- [7] Liu Q L, Gu X F, Li J P. Researches of Chinese sentence similarity based on HowNet[C]// International Conference on Apperceiving Computing & Intelligence Analysis. Chengdu, China, 2010, 26-29.
- [8] Dinu G, Thater S. Saarland: vector-based models of semantic textual similarity[C]// Joint Conference on Lexical & Computational Semantics. [S.l.]: 2012, 603-607.
- [9] 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.  
LI Bin, LIU Ting, QIN Bing, et al. Chinese sentence similarity computing based on semantic dependency relationship analysis[J]. Application Research of Computers, 2003, 20(12): 15-17. (in Chinese)
- [10] Oliva J, Serrano J I, Castillo M D D, et al. SyMSS: A syntax-based measure for short-text semantic similarity[J]. Data & Knowledge Engineering, 2011, 70(4): 390-405.
- [11] Chang J W, Lee M C, Wang T I, et al. Using grammar patterns to evaluate semantic similarity for short texts[C]// International Conference on Computing Technology and Information Management (NCM and ICNIT). [S. l.]: IEEE, 2012.
- [12] Šarić F, Glavaš G, Karan M, et al. Takelab: systems for measuring semantic text similarity[C]// Conference on Lexical and Computational Semantics. Montréal, Canada: 2012: 441-448.
- [13] 熊大平,王健,林鸿飞.一种基于 LDA 的社区问答问句相似度计算方法[J].中文信息学报,2012,26(5):40-46.  
XIONG Daping, WANG Jian, LIN Hongfei. An LDA-based approach to finding similar questions for community question answer[J]. Journal of Chinese Information Processing, 2012, 26(5): 40-46. (in Chinese)
- [14] 田卫东,强继朋.基于问句类型的问句相似度计算[J].计算机应用研究,2014,31(4):1090-1093.  
TIAN Weidong, QIANG Jipeng. Questions similarity computation based on question classification[J]. Application Research of Computers, 2014, 31(4): 1090-1093. (in Chinese)
- [15] 杨萌,李培峰,朱巧明,等.一种基于 Tree-LSTM 的句子相似度计算方法[J].北京大学学报(自然科学版),2018,54(3):

481-486.

YANG Meng, LI Peifeng, ZHU Qiaoming, et al. An approach of sentence similarity on tree-LSTM[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2018, 54(3): 481-486. (in Chinese)

- [16] 李峰,侯加英,曾荣仁,等.融合词向量的多特征句子相似度计算方法研究[J].计算机科学与探索,2017,11(4):608-618.  
LI Feng, HOU Jiaying, ZENG Rongren, et al. Research on multi-feature sentence similarity computing method with word embedding[J]. Journal of Frontiers of Computer Science & Technology, 2017, 11(4): 608-618. (in Chinese)
- [17] 邓涵,朱新华,李奇,等.基于句法结构与修饰词的句子相似度计算[J].计算机工程,2017,43(9):240-244,249.  
DENG Han, ZHU Xinhua, LI Qi, et al. Sentence similarity calculation based on syntactic structure and modifier[J]. Computer Engineering, 2017, 43(9): 240-244,249. (in Chinese)
- [18] 杜玉锋,季铎,姜利雪,等.基于 SAO 的专利结构化相似度计算方法[J].中文信息学报,2016,30(1):30-35.  
DU Yufeng, JI Duo, JIANG Lixue, et al. Patent similarity measure based on SAO structure[J]. Journal of Chinese Information Processing, 2016, 30(1): 30-35. (in Chinese)
- [19] 张绍阳,曹家波,王子凡,等.基于加权二部图匹配的中文段落相似度计算[J].计算机工程与应用,2017,53(18):95-101.  
ZHANG Shaoyang, CAO Jiabo, WANG Zifan, et al. Chinese paragraph similarity calculated based on weighted bipartite graph match[J]. Computer Engineering and Applications, 2017, 53(18): 95-101. (in Chinese)
- [20] 熊晶,王继鹏,魏墨济,等.基于细粒度依存关系的中文长句相似度计算[J].科学技术与工程,2017,17(11):277-281.  
XIONG Jing, WANG Jipeng, WEI Moji, et al. Chinese long sentences similarity calculation based on fine-grained dependency syntax[J]. Science Technology and Engineering, 2017, 17(11): 277-281. (in Chinese)
- [21] 黄洪,陈德锐.基于语义依存的汉语句子相似度改进算法[J].浙江工业大学学报,2017,45(1):6-9.  
HUANG Hong, CHEN Derui. An improved Chinese sentence similarity algorithm based on semantic dependency[J]. Journal of Zhejiang University of Technology, 2017, 45(1): 6-9. (in Chinese)
- [22] 蒲进兴,李登辉,李征,等.基于动态相似度的错误定位优先排序方法[J].计算机科学,2016,43(3):206-212.  
PU Jinxing, LI Denghui, LI Zheng, et al. Dynamic similarity based fault localization prioritization[J]. Computer Science, 2016, 43(3): 206-212. (in Chinese)

(编辑 詹燕平)