


中文信息相似度计算 理论与方法

夏 天 [著]

 河南科学技术出版社

中文信息相似度计算理论与方法

夏天 [著]

附图书引用方式:

夏天. 中文信息相似度计算理论与方法 [M]. 郑州: 河南科学技术出版社. 2009. ISBN: 978-7-5349-4434-5.

如需购买图书, 请大家直接在淘宝上搜索相似度计算, 掌柜名称为 xiatyi。

目 录

第 1 章	引论.....	7
1.1	背景及意义.....	7
1.2	研究现状.....	9
1.2.1	词语相似度计算的研究进展.....	10
1.2.2	组块相似度计算的研究进展.....	12
1.2.3	句子相似度计算的研究进展.....	12
1.2.4	文本相似度计算的研究进展.....	13
1.3	参考文献.....	14
第 2 章	相似度计算基础资源.....	17
2.1	自动分词程序资源.....	17
2.1.1	分词的必要性.....	17
2.1.2	简单分词系统的设计与实现.....	18
2.1.3	ICTCLAS 词法分析系统的的 JNI 调用.....	20
2.1.4	Java 分词系统 ICTCLAS4J.....	21
2.2	词典与语料库电子资源.....	23
2.2.1	同义词词林电子资源.....	24
2.2.2	知网电子资源.....	26
2.2.3	WordNet 电子资源.....	30
2.2.4	LCMC 汉语平衡语料库电子资源.....	32
2.2.5	人民日报标注语料库电子资源.....	34
2.3	本章小结.....	35
2.4	参考文献.....	35
第 3 章	汉语词语相似度计算.....	37
3.1	词语相似度的基本概念.....	37
3.1.1	词语相似度的定义与特点.....	37
3.1.2	词语相似度与相关性.....	38
3.2	主流词语相似度计算方法与特点.....	38
3.2.1	字面相似度计算方法及特点.....	38
3.2.2	词素相似度计算方法及特点.....	40
3.2.3	词林相似度计算方法及特点.....	42
3.2.4	知网相似度计算方法及特点.....	45
3.2.5	本体概念相似度计算及特点.....	47
3.2.6	统计相似度计算方法及特点.....	48
3.3	可扩展的词语相似度计算方法.....	48
3.3.1	未登录概念词的分布统计.....	48
3.3.2	义原相似度计算.....	49
3.3.3	概念词语的相似度计算.....	51
3.3.4	未登录概念词语的相似度计算.....	52
3.3.5	实验结果.....	55
3.3.6	对建设概念词典的启示.....	57
3.4	本章小结.....	58

3.5	参考文献.....	58
第 4 章	汉语组块相似度计算.....	60
4.1	引言.....	60
4.2	组块基本概念.....	60
4.2.1	组块的定义和分类.....	60
4.2.2	汉语组块的自动识别.....	62
4.3	组块与短语级别的相似度计算现状.....	64
4.4	组块相似度计算的基本要求.....	66
4.5	组块相似度计算方法.....	66
4.5.1	极大筛选法.....	66
4.5.2	最大匹配均值法.....	67
4.5.3	权重动态分配法.....	67
4.5.4	权重动态分配法计算过程.....	69
4.5.5	实验结果与分析.....	70
4.6	本章小结.....	70
4.7	参考文献.....	71
第 5 章	汉语句子相似度计算.....	72
5.1	引言.....	72
5.2	句子相似度定义与分类.....	72
5.3	汉语句子相似度计算的特点与难点.....	73
5.4	句子相似度计算的常用方法.....	74
5.4.1	以 TF/IDF 为代表的统计方法.....	74
5.4.2	语义词典方法.....	75
5.4.3	词形与词序综合法.....	75
5.4.4	依存树法.....	76
5.4.5	编辑距离法.....	77
5.5	改进编辑距离算法与句子相似度计算.....	77
5.5.1	编辑距离算法原理.....	77
5.5.2	标准编辑距离算法实现.....	78
5.5.3	编辑单元的重新确定.....	80
5.5.4	面向语义的代价函数.....	80
5.5.5	支持非相邻块交换的编辑距离算法.....	85
5.5.6	算法优化.....	90
5.5.7	归一化与句子相似度计算.....	92
5.5.8	算法实现.....	92
5.5.9	实验结果与分析.....	92
5.6	本章小结.....	94
5.7	参考文献.....	94
第 6 章	文本相似度计算.....	96
6.1	引言.....	96
6.1.1	文本内容相似度和结构相似度.....	96
6.1.2	文本相似度计算方法分类.....	96
6.2	文本内容的数学描述.....	97
6.2.1	经典文本计算模型.....	97

6.2.2	文本表示方法.....	102
6.3	文本内容相似度计算方法.....	104
6.3.1	基于散列函数的相似度计算.....	104
6.3.2	基于距离的相似度计算.....	105
6.3.3	基于集合理论的相似度计算.....	106
6.3.4	基于向量的相似度计算.....	108
6.3.5	基于语义的文本相似度计算.....	116
6.4	文本结构相似度计算.....	117
6.4.1	基本概念.....	117
6.4.2	基于标记的结构相似度计算.....	120
6.4.3	基于 Path Shingle 的结构相似度计算.....	120
6.4.4	基于树编辑距离算法的结构相似度计算.....	123
6.5	本章小结.....	128
6.6	参考文献.....	129
第 7 章	相似度计算应用篇.....	131
7.1	基于 ALICE 的汉语自然语言接口.....	131
7.1.1	ALICE 简介.....	131
7.1.2	人工智能标记语言 AIML.....	132
7.1.3	基于 ALICE 的汉语自然语言接口 CNLIS.....	135
7.1.4	CNLIS 系统结构.....	137
7.1.5	实验结果与分析.....	139
7.2	面向 FAQ 的问答系统 FAQ-Guider.....	140
7.2.1	FAQ 问答简介.....	140
7.2.2	设计思想.....	140
7.2.3	系统实现.....	141
7.2.4	实验结果与分析.....	144
7.3	Web 文档的结构聚类.....	145
7.3.1	Shingle 距离矩阵和 Web 文档相似度.....	145
7.3.2	结构聚类算法.....	146
7.3.3	实验结果.....	147
7.4	本章小结.....	148
7.5	参考文献.....	148
第 8 章	结论与展望.....	150
8.1	结论.....	150
8.2	展望.....	151
附 1:	代码获取说明.....	152