第 31 卷 第 2 期　　华 中 科 技 大 学 学 报(自然科学版)　　Vol. 31 No. 2
2003 年 2 月　　J. Huazhong Univ. of Sci. & Tech. (Nature Science Edition)　　Feb. 2003

基于语义的模糊匹配探索与应用

程 莉　卢正鼎　文坤梅　李 娟

(华中科技大学计算机科学与技术学院)

摘要:提出计算词与词之间的相似度,通过比较词语相似度来确定搜索到的信息是否属于某一类特定信息,从而实现基于语义的模糊匹配.实验结果表明,该方法较传统的精确匹配方法、模糊串匹配方法能更好地保存有用信息,提高了过滤结果的完整性和准确性.

关 键 词:语义计算;相似度;模糊匹配;知网

中图分类号:TP301.6　　**文献标识码**:A　　**文章编号**:1671-4512(2003)02-0023-03

1 语义计算过程

在知网中义原(primitive)的关系是通过几个特征文件建立起来的,每个义原之间又存在复杂的关系,如上下位关系、同义关系、反义关系等等.所以,义原之间组成的是一个复杂的网状结构,而不是一个单纯的树状结构.不过,义原关系中最重要的还是上下位关系.根据义原的上下位关系,所有的'基本义原'组成了一个义原层次体系(如图1).这个义原层次体系是一个树状结构,这也是进行语义相似度计算的基础.

-entity|实体
　├thing|万物
　… ├physicsal|物质
　　… ├nimate|生物
　　　… ├Ani8malHuman|动物
　　　　… ├human|人
　　　　| └humanized|拟人
　　　　└animal|兽
　　　　　├beast|走兽
　　　　　…

图 1　树状的义原次结构

在知网中,每一个概念是通过一组义原来表示的,每个记录的具体记录格式如下:

NO. = 词或短语序号;
[W_X = 词或短语;
G_X = 词或短语的词性;
E_X = 词或短语的例子] + ;
DEF = 概念定义,

其中的 W_X、G_X、E_X 构成每种语言的记录,X 用以描述记录所代表语种,X 为 C 则为汉语,为 E 则为英语.每个词语由 DEF 来描述其概念定义,DEF 的值由若干个义原及它们与主干词之间的语义关系描述组成.

概念(具体词)本身并不是义原层次体系中的一个结点,义原才是这个层次体系中的一个结点.而且,一个概念并不是简单地描述为义原的集合,而是要描述为使用某种专门的'知识描述语言'来表达的一个语义表达式.也就是说,在描述一个概念的多个义原中,每个义原所起到的作用是不同的,这就给我们的相似度计算带来了很大的困难.

1.1 义原相似度计算

由于所有的概念都最终归结于用义原(个别地方用具体词)来表示,所以义原的相似度计算是概念相似度计算的基础.

定义 义原 p_1 和 p_2 的语义相似度:

$$\mathrm{sim}(p_1, p_2) = \alpha/(d + \alpha), \qquad (1)$$

式中,p_1 和 p_2 表示两个义原;d 是 p_1 和 p_2 在义原层次体系中的路径长度,是一个正整数;α 是一个可调节的参数,α 的含义是当相似度为 0.5 时的词语距离值.为了简单起见,这里主要利用了义原的上下位关系.

显然,只有当两个义原处于一个分类体系中,才具有相似度,且相似度和义原之间的距离成反比.

因为在知网的知识描述语言中,在一些义原

收稿日期:2002-08-15.
作者简介:程 莉(1979-),女,硕士研究生 武汉,华中科技大学计算机科学与技术学院 (430074).
基金项目:国家高性能计算基金资助项目(99319).

出现的位置可能出现一个具体词(概念),并用圆括号()括起来.为了简化起见,规定:

a. 具体词与义原的相似度一律处理为一个比较小的常数(γ);

b. 具体词和具体词的相似度,如果两个词相同,则为 1,否则为 0.

1.2 实词概念的相似度计算

本文的研究主要用以提高匹配有效信息的准确率,所以词语相似度主要指实词的相似度.即两个词语在不同的上下文中可以互相替换使用而不改变文本的句法语义结构的程度.两个词语互换的可能性越大,二者的相似度就越高,否则相似度就越低.

基于知网对词语义项的概念及其描述格式,作如下定义:两个概念语义表达式的整体相似度记为

$$\mathrm{sim}(S_1,S_2)=\sum_{i=1}^{4}\beta_i\mathrm{sim}_i(S_1,S_2), \quad (2)$$

式中,β_i $(1\leqslant i\leqslant 4)$ 是可调节的参数,且有 $\beta_1+\beta_2+\beta_3+\beta_4=1$,$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$;$\mathrm{sim}_1(S_1,S_2)$ 表示第一独立义原描述式的相似度,$\mathrm{sim}_2(S_1,S_2)$ 表示其他独立义原描述式的相似度,$\mathrm{sim}_3(S_1,S_2)$ 表示关系义原描述式的相似度,$\mathrm{sim}_4(S_1,S_2)$ 表示符号义原描述式的相似度.由于第一独立义原描述式反映了一个概念最主要的特征,因此应该将其权值定义得比较大,一般应在 0.5 以上,β_i 递减反映了 sim_1 到 sim_4 对于总体相似度所起到的作用依次递减.如果某一部分的对应物为空,则任何义原(或具体词)与空值的相似度定义为一个比较小的常数(δ).

在实验中我们发现,如果 sim_1 非常小,但 sim_3 或者 sim_4 比较大,将导致整体的相似度仍然比较大的不合理现象.因此对公式(2)进行了修改,得到

$$\mathrm{sim}(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}\mathrm{sim}_j(S_1,S_2). \quad (3)$$

其意义在于,主要部分的相似度值对于次要部分的相似度值起到制约作用,也就是说,如果主要部分相似度比较低,那么次要部分的相似度对于整体相似度所起到的作用也要降低.下面再分别讨论每一部分的相似度.

A. 第一独立义原描述式.就是两个义原的相似度,按照公式(1)计算即可.

B. 其他独立义原描述式.因为其他独立义原描述式不止一个,所以计算较为复杂.可把整体相似度还原为部分相似度的加权平均.困难在于,各个独立义原描述式之间没有分工,所以很难找到对应关系.按照如下步骤对这些独立义原描述式分组:

a. 先把两个表达式的所有独立义原(第一个除外)任意配对,计算出所有可能的配对的义原相似度;

b. 取相似度最大的一对,并将它们归为一组;

c. 在剩下的独立义原的配对相似度中,取最大的一对,并归为一组,如此反复,直到所有独立义原都完成分组.

C. 关系义原描述式.关系义原描述式的配对分组较为简单,把关系义原相同的描述式分为一组,并计算其相似度;

D 符号义原描述式.符号义原描述式的配对分组与关系义原描述式类似,把关系符号相同的描述式分为一组,并计算其相似度.

E. 在以上 B~D 的计算中,最后求加权平均时,各部分取相等的权值.

1.3 词语间相似度的计算

定义 两个汉语词语 W_1 和 W_2,如果 W_1 有 n 个义项(概念):S_{11}, S_{12}, …, S_{1n}, W_2 有 m 个义项(概念):S_{21}, S_{22}, …, S_{2m},则规定,W_1 和 W_2 的相似度各个概念的相似度之最大值

$$\mathrm{sim}(W_1,W_2)=\max_{i=1\cdots n,j=1\cdots m}\mathrm{sim}(S_{1i},S_{2j}). \quad (4)$$

这样,就把两个词语之间的相似度问题归结到了两个概念之间的相似度问题.至此,可以准确方便地利用知网计算出基于语义的两词语间相似度.

2 实验结果分析和讨论

实验主要比较精确匹配、模糊串匹配和本文提出的基于语义模糊匹配三种方法对词语匹配的效果.在计算语义相似度时,参数取值如下:$\alpha=1.6$;$\beta_1=0.5$,$\beta_2=0.2$,$\beta_3=0.17$,$\beta_4=0.13$;$\gamma=0.2$,γ 为具体词与义原的相似度;$\delta=0.2$,为义原(或词)与定值的似度.

计算出词语间相似度后,可根据需要设定阈值.本试验规定:当两词语的相似度$\geqslant$0.6 时,即认为他们属于可替换的近义词.由于条件限制,对于匹配结果采用了人工判别的方法,即由人来判断这个词和该组词的相似度大小是否与人的直觉相符合.实验结果如表 1.

表 1　实验结果

词语 1	词语 2	精确匹配	模糊串匹配	语义模糊匹配
精巧	精巧	√	√	√
精巧	精致	×	√	√
精巧	别致	×	×	√
佳肴	佳肴	√	√	√
佳肴	菜肴	×	√	√
佳肴	美味	×	×	√
夸大	夸大	√	√	√
夸大	夸张	×	√	√
夸大	扩充	×	×	√
安全	可靠	×	×	√
疗效显著	效果明显	×	×	√
国家	政府	×	×	√
男人	和尚	×	×	√
女人	母亲	×	×	√
高效	快速	×	×	√
同意	允许	×	×	√

从实验结果可以看出：

a. 采用基于语义的模糊匹配，匹配成功率远远高于传统的精确匹配和字符串模糊匹配；

b. 由于知网采用义原，通过一种知识描述语言来对每个概念进行描述，可能导致两词语虽然相似度高，但在具体的语言环境中仍不适合替换. 即匹配准确率有待提高.

参　考　文　献

[1] Gan K W, Wong P W. Annotating information structures in Chinese texts using HowNet. Hong Kong: Second Chinese Language Processing Workshop, 2000. 85～92

[2] 李素建. 基于语义计算的语句相关度研究. 计算机工程与应用, 2002 (7): 75～76

The exploration of amphibolous matching based on semantics and its application

Cheng Li　Lu Zhendin　Wen Kunmei　Li Juan

Abstract: Semantic computation was introduced. The information could be judged if needed by computing the similarity between words. With the help of this way the amphibolous matching based on semantics was realized. The experiment result shows that our method is effective. Compared with the traditional methods, it has improved the integrity and veracity of filtering results.

Key words: semantic computation; similarity; amphibolous matching; HowNet

Cheng Li　Postgraduate; College of Computer Sci. & Tech., Huazhong Univ. of Sci. & Tech., Wuhan 430074, China.

(上接第 8 页)

The design of hot-backup mechanism with multi-server and its implementation

Feng Yucai, Wang Dongmin, Zhu Hong

Abstract: A backup technologies in moment use was analyzed. Hot-backup mechanism with multi-server was designed, and a voting arithmetic was given. The multi-application and multi-database servers were operated cooperatively, and the stabilizability and credibility of the application system ensured. This system has been used to the Wuhan Fire Communication and Command System. The stability and reliability of the system are proved.

Key words: data base; multi-server; hot-backup; servers switch

Feng Yucai　Prof.; College of Computer Sci. & Tech., Huazhong Univ. of Sci. & Tech., Wuhan 430074, China.

基于语义的模糊匹配探索与应用

作者： 程莉， 卢正鼎， 文坤梅， 李娟
作者单位： 华中科技大学计算机科学与技术学院
刊名： 华中科技大学学报(自然科学版) ISTIC EI PKU
英文刊名： JOURNAL OF HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY(NATURE SCIENCE)
年，卷(期)： 2003,31(2)
被引用次数： 15次

参考文献(2条)

1. 李素建 基于语义计算的语句相关度研究[期刊论文]-计算机工程与应用 2002(07)

2. Gan K W;Wong P W Annotating information structures in Chinese texts using HowNet 2000

引证文献(15条)

1. 李佳林 在线考试系统中主观题自动阅卷的设计[期刊论文]-中国教育技术装备 2008(24)

2. 文坤梅. 卢正鼎. 叶卫国 Web-MIND:基于特定主题的Web信息挖掘系统[期刊论文]-计算机工程与科学 2007(6)

3. 张茂元. 邹春燕. 卢正鼎 一种基于语义匹配的Web信息提取方法研究[期刊论文]-计算机工程与应用 2006(23)

4. 刘亚清. 于纯妍. 张瑾 改进的基于义原同现频率的汉语词义排歧方法[期刊论文]-计算机工程与科学 2006(12)

5. 张瑾. 刘亚清. 于纯妍 汉语词义排歧的另一种方法[期刊论文]-小型微型计算机系统 2006(4)

6. 刘亚清. 张瑾. 于纯妍 基于义原同现频率的汉语词义排歧系统[期刊论文]-计算机技术与发展 2006(5)

7. 朱雪刚 基于语义网络的教学资源搜索引擎研究[学位论文]硕士 2006

8. 汪泱 主观文字试题评判相关技术研究[学位论文]硕士 2006

9. 刘子君 XML树状结构的数据挖掘结构相似性算法[学位论文]硕士 2006

10. 卢正鼎. 张茂元 一种基于义素的网页信息项语义匹配方法研究[期刊论文]-计算机科学 2005(4)

11. 许超 汉英双语网页资源中相同事件文本对的提取[学位论文]硕士 2005

12. 倪应华 基于XML在线考试系统的研究与实现[学位论文]硕士 2005

13. 骆婕 基于语义的图形检索的研究与实现[学位论文]硕士 2005

14. 周新栋 中文文本分类的文档索引机制及分类模型的研究[学位论文]硕士 2004

15. 丁一 基于Web挖掘的个性化推荐服务研究[学位论文]硕士 2004