基于语义的协作信息检索框架研究

报告人： 盛艳
导 师： 李云

扬州大学信息工程学院

2009 年 04 月 20 日

Outline

研究背景

- Web信息与日俱增，其种类结构也是丰富多样，尤其文本信息更是巨大，为了能在大数据集中更快速，更有效的检索出更符合用户需求的结果，使用基于语义的一系列策略.

研究背景

- Web信息与日俱增，其种类结构也是丰富多样，尤其文本信息更是巨大，为了能在大数据集中更快速，更有效的检索出更符合用户需求的结果，使用基于语义的一系列策略.

- 提取大数据集中有意义的特征;

研究背景

- Web信息与日俱增，其种类结构也是丰富多样，尤其文本信息更是巨大，为了能在大数据集中更快速，更有效的检索出更符合用户需求的结果，使用基于语义的一系列策略.

- 提取大数据集中有意义的特征；
- 基于概念格的理论和技术，构建多个概念格，并根据概念间的相似度进行匹配，并可在分布式环境下进行检索；

研究背景

- Web信息与日俱增，其种类结构也是丰富多样，尤其文本信息更是巨大，为了能在大数据集中更快速，更有效的检索出更符合用户需求的结果，使用基于语义的一系列策略.

- 提取大数据集中有意义的特征；
- 基于概念格的理论和技术，构建多个概念格，并根据概念间的相似度进行匹配，并可在分布式环境下进行检索；
- 通过对用户的历史查询记录进行提取，建立用户兴趣模型并利用该模型对检索结果重新排序；

研究背景

- Web信息与日俱增，其种类结构也是丰富多样，尤其文本信息更是巨大，为了能在大数据集中更快速，更有效的检索出更符合用户需求的结果，使用基于语义的一系列策略.

- 提取大数据集中有意义的特征;
- 基于概念格的理论和技术，构建多个概念格，并根据概念间的相似度进行匹配，并可在分布式环境下进行检索;
- 通过对用户的历史查询记录进行提取，建立用户兴趣模型并利用该模型对检索结果重新排序;
- 这样形成了一整套更具有时空效率又能符合用户个性的信息检索方案.

国内外研究现状

- 概念作为人的思想和知识的基本单元，一直以来，深受哲学界和科学界的重视，很自然地，也就成为了人工智能学科的重要研究对象，这主要体现在知识表示和机器学习等领域.将概念格理论应用到信息检索中，更能提取出概念之间的本质联系，对检索准确度有一个更深层次的提高.

国内外研究现状

- 概念作为人的思想和知识的基本单元，一直以来，深受哲学界和科学界的重视，很自然地，也就成为了人工智能学科的重要研究对象，这主要体现在知识表示和机器学习等领域.将概念格理论应用到信息检索中，更能提取出概念之间的本质联系，对检索准确度有一个更深层次的提高.

- 概念格是对概念以及概念之间关系的描述，是形式概念分析(Formal Concept Analysis, FCA)的核心数据结构.它在一定程度上是对客观世界的一种高度简化的描述形式.这种简化的最大优点是其具有良好的数学性质.Wille基于这种简化，开创了形式概念分析领域.但是，正是由于这种简化，使得概念格不能简单地被理解为客观世界的部分模型，而更多的是被看作一个人造的从一些数据集派生出来的表示.

国内外研究现状

- 国内外针对信息检索领域中提出了很多方法及理论，其中，使用形式概念分析理论来进行信息检索也有很多深入的研究，如，使用FCA构建表示信息模型以更确切的表示原始检索模型，并使用构建出来的概念格进行信息检索，基于抽象具体程度的不同进行检索可获得独特层面上的检索结果.

国内外研究现状

- 国内外针对信息检索领域中提出了很多方法及理论，其中，使用形式概念分析理论来进行信息检索也有很多深入的研究，如，使用FCA构建表示信息模型以更确切的表示原始检索模型，并使用构建出来的概念格进行信息检索，基于抽象具体程度的不同进行检索可获得独特层面上的检索结果.

- 另外，在分布式环境下，使用FCA理论和技术进行协作信息检索，这可大大提高检索时间和效率.

论文研究目标

- 针对文本中的特征提取和约简方法.

论文研究目标

- 针对文本中的特征提取和约简方法.
- 概念之间的结构和语义相似度度量方法.

论文研究目标

- 针对文本中的特征提取和约简方法.
- 概念之间的结构和语义相似度度量方法.
- 协作信息检索系统的分治合并策略.

论文研究目标

- 针对文本中的特征提取和约简方法.
- 概念之间的结构和语义相似度度量方法.
- 协作信息检索系统的分治合并策略.
- 用户兴趣模型的构建方法及个性化检索模型.

论文主要内容

- 根据文本各个特征自身和类别信息的统计特性，提取更少的特征来尽可能的表达出文本蕴含的信息，得到一种有效的文本特征提取方法，其中也对文本特征进行约简从而得到少而精的特征信息.

论文主要内容

- 根据文本各个特征自身和类别信息的统计特性，提取更少的特征来尽可能的表达出文本蕴含的信息，得到一种有效的文本特征提取方法，其中也对文本特征进行约简从而得到少而精的特征信息.
- 基于概念格理论，结合粗糙集理论和语义本体理论，从形式概念的结构相似和语义相似两个层次上衡量概念间的相似度形式一种有效确切的概念相似度衡量方法.

论文主要内容

- 根据文本各个特征自身和类别信息的统计特性，提取更少的特征来尽可能的表达出文本蕴含的信息，得到一种有效的文本特征提取方法，其中也对文本特征进行约简从而得到少而精的特征信息.
- 基于概念格理论，结合粗糙集理论和语义本体理论，从形式概念的结构相似和语义相似两个层次上衡量概念间的相似度形式一种有效确切的概念相似度衡量方法.
- 针对分布式环境，提出并构建一个协作信息检索框架，实现从结构和语义两个层次上的概念匹配，获得更符合用户需求的检索结果.

论文主要内容

- 根据文本各个特征自身和类别信息的统计特性，提取更少的特征来尽可能的表达出文本蕴含的信息，得到一种有效的文本特征提取方法，其中也对文本特征进行约简从而得到少而精的特征信息.
- 基于概念格理论，结合粗糙集理论和语义本体理论，从形式概念的结构相似和语义相似两个层次上衡量概念间的相似度形式一种有效确切的概念相似度衡量方法.
- 针对分布式环境，提出并构建一个协作信息检索框架，实现从结构和语义两个层次上的概念匹配，获得更符合用户需求的检索结果.
- 在对检索结果排序过程中，利用语义本体WordNet对用户历史检索记录进行分析，从而构建用户兴趣模型，以便更精确的表达出用户兴趣所在，并按照此模型对初始检索结果重新排序.

采取的研究方案及可行性分析

基于类别分析及有效特征提取的文本分类方法

- 结合人类语言中普遍存在的Zipf定律，对大量文本集进行全局的特征提取.

$$P(r) = \frac{C}{r^{\alpha}}$$

采取的研究方案及可行性分析

基于类别分析及有效特征提取的文本分类方法

- 结合人类语言中普遍存在的Zipf定律，对大量文本集进行全局的特征提取.

$$P(r) = \frac{C}{r^{\alpha}}$$

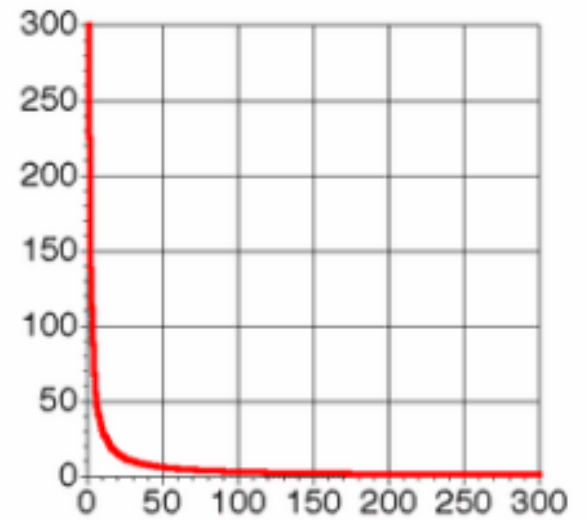

Figure 1 Zipf Law's frequency-ranking curve

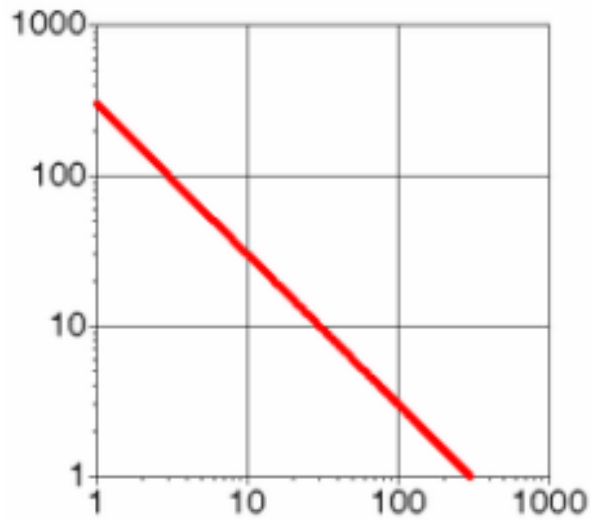

Figure 2 Zipf Law's log-log curve

采取的研究方案及可行性分析

基于类别分析及有效特征提取的文本分类方法

- 在计算每个特征词的类别频率CF之后，对文本集进行类内局部特征提取.
  CF计算公式:

$$CF(t, C_i) = \frac{the\ times\ of\ t\ appears\ in\ c_i}{the\ number\ of\ all\ words\ in\ c_i}$$

采取的研究方案及可行性分析

基于类别分析及有效特征提取的文本分类方法

- 在计算每个特征词的类别频率CF之后，对文本集进行类内局部特征提取.
  CF计算公式:

$$CF(t, C_i) = \frac{the\ times\ of\ t\ appears\ in\ c_i}{the\ number\ of\ all\ words\ in\ c_i}$$

- 在CF基础上，提出一个特征权值公式.

$$\omega(t,d) = \frac{TF(t,d) \times \log(IDF(t) + \sigma) \times max_{c \in C} CF(t,c)}{\sqrt{\sum_{t \in d}(TF(t,d) \times \log(IDF(t) + \sigma) \times max_{c \in C} CF(t,c))^2}}$$

采取的研究方案及可行性分析

基于类别分析及有效特征提取的文本分类方法

- 最后，使用提出的权值公式，改进K近邻算法，对标准数据集NewsGroup进行文本自动分类测试，得到效果如下图：

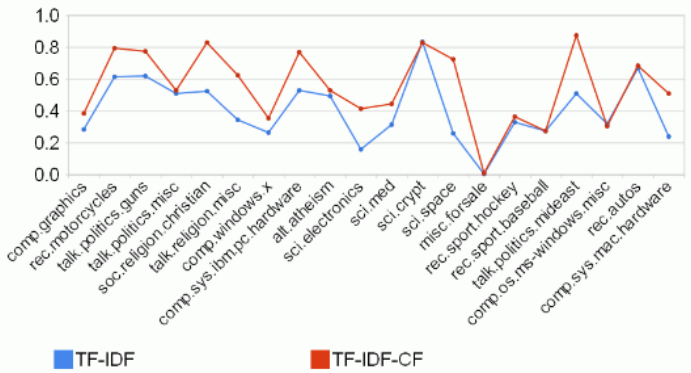

Figure 3 the result of recall(r)

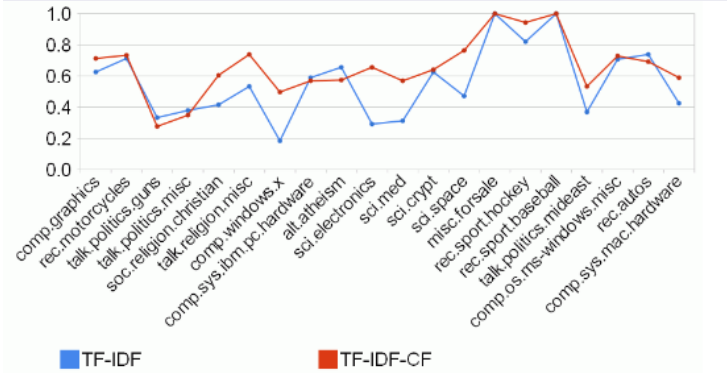

Figure 4 the result of precision(p)

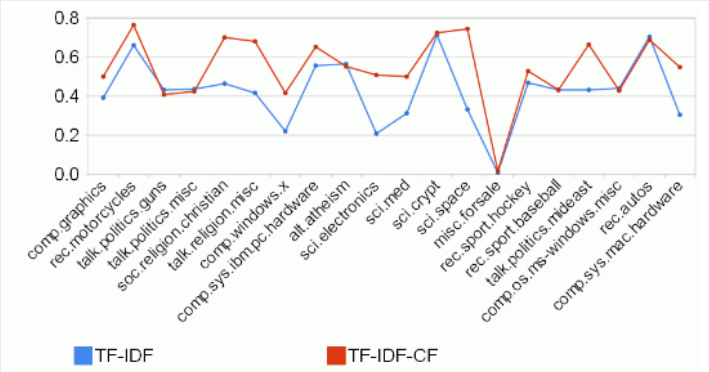

Figure 5 the result of F-guideline

采取的研究方案及可行性分析

结构及语义层次上的概念相似度衡量方法

- 针对文本特征的形式背景， 按照已经成熟的概念格构造算法， 使用粗糙集中的下近似并结合概念的层次信息， 语义本体的语义信息衡量概念间的相似度并实际应用到协作信息检索过程中.

采取的研究方案及可行性分析

结构及语义层次上的概念相似度衡量方法

- 针对文本特征的形式背景， 按照已经成熟的概念格构造算法， 使用粗糙集中的下近似并结合概念的层次信息， 语义本体的语义信息衡量概念间的相似度并实际应用到协作信息检索过程中.
- 结构相似度计算方法

$$StructSim(C_1, C_2) = \frac{|(B_1 \cap B_2)_{LA}|}{|(B_1 \cup B_2)_{LA}| + \alpha(C_1, C_2)|B_{1LA} - B_{2LA}| + (1 - \alpha(C_1, C_2))|B_{2LA} - B_{1LA}|}$$

采取的研究方案及可行性分析

结构及语义层次上的概念相似度衡量方法

- 针对文本特征的形式背景，按照已经成熟的概念格构造算法，使用粗糙集中的下近似并结合概念的层次信息，语义本体的语义信息衡量概念间的相似度并实际应用到协作信息检索过程中.
- 结构相似度计算方法

$$StructSim(C_1, C_2) = \frac{|(B_1 \cap B_2)_{LA}|}{|(B_1 \cup B_2)_{LA}| + \alpha(C_1, C_2)|B_{1LA} - B_{2LA}| + (1 - \alpha(C_1, C_2))|B_{2LA} - B_{1LA}|}$$

- 语义相似度的计算是通过Wordnet找到词之间的相互关系，通过一定策略进行量化.

$$SemanticSim(C_1, C_2) = \frac{M(B_1, B_2)}{max\{|B_1|, |B_2|\}}$$

采取的研究方案及可行性分析

一种基于WordNet的检索结果个性化排序方法

- 在对检索结果排序过程中，利用语义本体WordNet对用户历史检索记录进行分析，从而构建用户兴趣模型，以便更精确的表达出用户兴趣所在，并按照此模型对初始检索结果重新排序。此用户模型是通过寻找各个历史关键词之间的联系进行构建，最终构建的如下图所示：

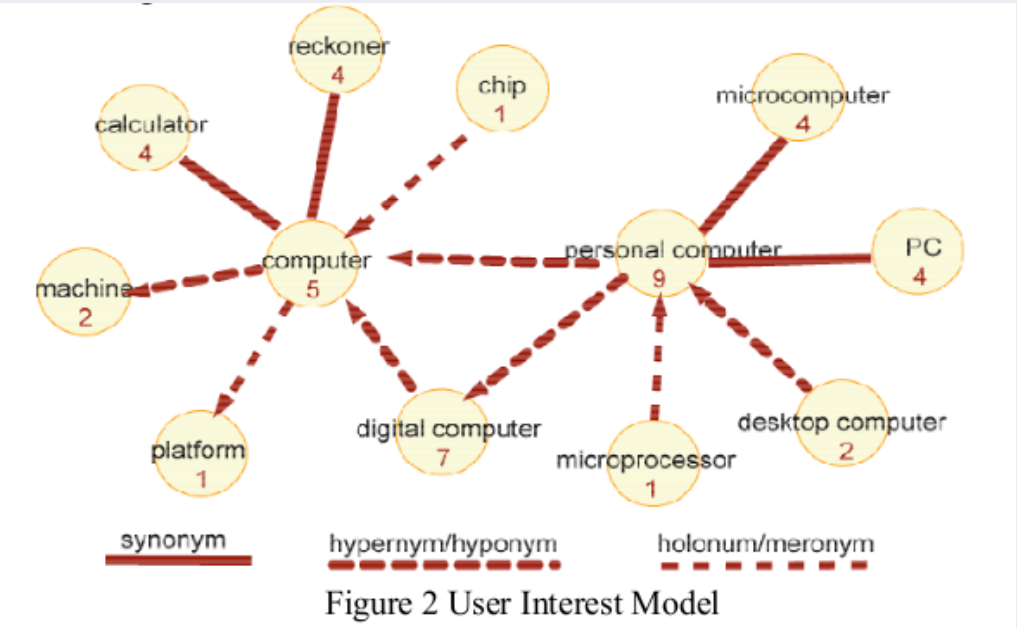

Figure 2 User Interest Model

论文技术路线

基于FCA的协作信息检索框架

- 分布式环境下，采用先分后合策略，对多个独立的形式概念格进行概念匹配后合并为最终结果返回给用户.

论文技术路线

基于FCA的协作信息检索框架

- 分布式环境下，采用先分后合策略，对多个独立的形式概念格进行概念匹配后合并为最终结果返回给用户.整体框图如下所示：

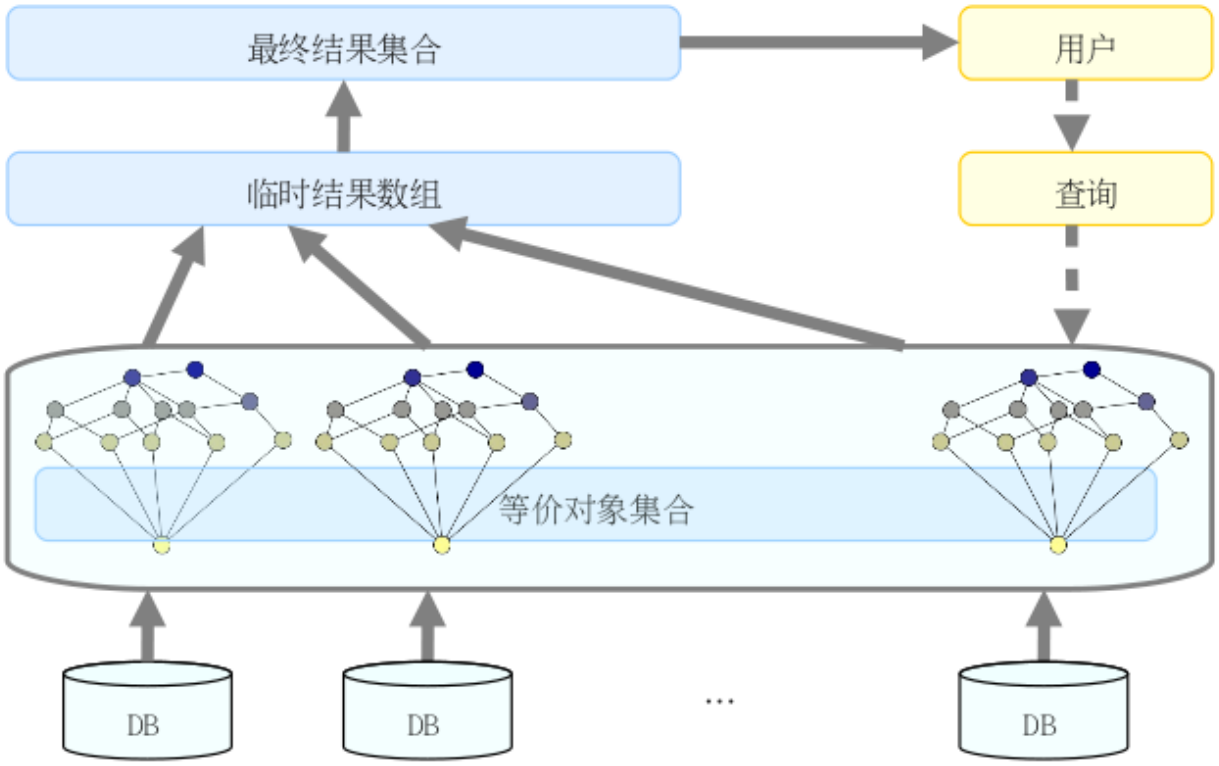

图 4 协作概念信息检索系统框图

论文技术路线

总结

- 用具体实例验证所提模型，方法和算法.
- FCA，文本特征提取策略，语义本体，粗糙集的理论提供了思想，方法和算法的研究基础.

论文后续扩展

应用方向

- 协作信息检索性能提高
- 利用本体的用户兴趣模型更新及维护
- 扩展到中文领域

理论方向

- 语义相似的进一步研究

参与的科研项目

- 面向本体的形式概念分析扩展模型及算法(60575035)，国家自然科学基金项目
- 序列概念格扩展模型及其序列模式挖掘算法研究(08KJB520012)，江苏省教育厅高校自然科学研究计划项目
- 基于概念格模型的本体构建与运算，扬州大学信息工程学院研究生创新基金项目

发表论文

- 一种基于概念格的本体合并方法，微电子学与计算机，Vol.25，NO.9，2008:34-36
- 基于概念格模型的本体映射，南京师范大学学报(工程技术版)，Vol.8，NO.4，2008:91-94，获第三届江苏计算机大会(JSCC2008)优秀论文
- A Rough Concept Lattice Model of Variable Precision，Proc. of IITA 2008 :162-166
- 基于概念格分布并行处理框架的知识发现研究，扬州大学博士启动基金项目
- 基于概念格模型的本体构建与运算，扬州大学信息工程学院研究生创新基金项目
- 多关系频繁项集的并行获取，微电子学与计算机，Vol.25，NO.10，2008:94-96
- A Text Classification Method with an Effective Feature Extraction based on Category Analysis(已投)
- A Personalized Search Results Ranking Method Based on WordNet (已投)

谢谢大家！