



(12) 发明专利申请

(10) 申请公布号 CN 102063497 A

(43) 申请公布日 2011. 05. 18

(21) 申请号 201010619675. 6

(22) 申请日 2010. 12. 31

(71) 申请人 百度在线网络技术(北京)有限公司

地址 100085 北京市海淀区上地十街 10 号
百度大厦

(72) 发明人 邓亮 陈浩然 来瑾颖 唐益龙
梁东杰 耿磊 李永强 严冰
韦晨曦 乔桥

(74) 专利代理机构 深圳市威世博知识产权代理
事务所(普通合伙) 44280

代理人 何青瓦 李庆波

(51) Int. Cl.

G06F 17/30 (2006. 01)

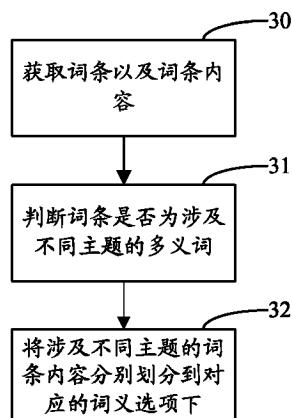
权利要求书 4 页 说明书 12 页 附图 5 页

(54) 发明名称

一种开放式知识共享平台及其词条处理方法

(57) 摘要

本发明提供了一种开放式知识共享平台及其词条处理方法。该词条处理方法包括：获取词条及词条内容；判断词条是否为涉及不同主题的多义词；若词条为涉及不同主题的多义词，则将涉及不同主题的词条内容分别划分到对应的词义选项下。通过上述所提供的技术方案，本发明提供了一种开放式知识共享平台及其词条处理方法，可将涉及不同主题的词条内容划分到不同词义选项下，使得开放式知识共享平台内容的粒度更细，更便于对同名不同主题的内容进行编辑完善和有针对性的引入扩展内容，从而提升用户的浏览体验。



1. 一种开放式知识共享平台的词条处理方法,其特征在于,所述词条处理方法包括:
 - a. 获取词条及词条内容;
 - b. 判断所述词条是否为涉及不同主题的多义词;
 - c. 若所述词条为涉及不同主题的多义词,则将涉及不同主题的所述词条内容分别划分到对应的词义选项下。
2. 根据权利要求 1 所述的词条处理方法,其特征在于,所述词条内容包括多个目录,在所述步骤 b 中,根据所述目录中的目录信息判断所述词条是否为涉及不同主题的多义词。
3. 根据权利要求 1 所述的词条处理方法,其特征在于,在所述步骤 a 中,进一步获取所述词条的分类信息,在所述步骤 b 中,根据所述分类信息判断所述词条是否为涉及不同主题的多义词。
4. 根据权利要求 1 所述的词条处理方法,其特征在于,所述步骤 b 包括:
 - b1. 对所述词条内容进行特征提取,以获取多个词条特征;
 - b2. 获取标注特征集合,所述标注特征集合包括多个带有权重值的标注特征;
 - b3. 根据所述标注特征分别为每一所述词条特征分配对应的权重值;以及
 - b4. 对所述多个词条特征的权重值进行求和,并将权重值求和高于阈值的词条作为多义词。
5. 根据权利要求 4 所述的词条处理方法,其特征在于,所述步骤 b2 进一步包括:
 - b21. 获取包括多个多义词样本和非多义词样本的标注语料;
 - b22. 从所述标注语料中提取多个所述标注特征;
 - b23. 根据所述标注特征在所述多义词样本和所述非多义词样本的出现情况为所述标注特征分别分配对应的权重值。
6. 根据权利要求 5 所述的词条处理方法,其特征在于,在所述步骤 b23 中,计算所述标注特征在所述多义词样本中出现的次数 M,计算所述标注特征在所述非多义词样本中出现的次数 N,并计算所述权重值等于 $M / (M+N)$ 。
7. 根据权利要求 1 所述的词条处理方法,其特征在于,在所述步骤 c 中,以所述词条的一级目录以及所述一级目录下的目录内容作为基本预测单元,预测所述基本预测单元是否属于不同主题,并根据预测结果将属于不同主题的所述基本预测单元分别划分到对应的词义选项。
8. 根据权利要求 7 所述的词条处理方法,其特征在于,在所述步骤 c 中,根据所述一级目录中的目录信息进行预测。
9. 根据权利要求 7 所述的词条处理方法,其特征在于,所述步骤 c 进一步包括:
 - c1. 从所述基本预测单元中提取多个目录特征;
 - c2. 根据所述目录特征判断所述基本预测单元之间的关联度,并根据所述关联度产生所述预测结果。
10. 根据权利要求 1 所述的词条处理方法,其特征在于,所述步骤 c 进一步包括:
 - c1. 获取所述词条的目录信息;
 - c2. 根据所述目录信息提取目录特征;
 - c3. 获取含目录特征关联度关系的机器模型;
 - c4. 根据所提取的所述目录特征,应用所述机器模型对前后相邻的所述目录信息的所

述目录特征进行关联度计算；

c5. 根据关联度计算结果对所述目录信息进行标记。

11. 根据权利要求 10 所述的词条处理方法，其特征在于，所述步骤 c2 进一步包括：在提取所述目录特征时，先进行分词。

12. 根据权利要求 11 所述的标的词分类分级方法，其特征在于，所述分词的方法包括：正向匹配分词、反向匹配分词、正向反向匹配分词、基于全切分词图的分词、最大熵马尔科夫模型分词、最大熵分词或条件随机场分词。

13. 根据权利要求 10 所述的词条处理方法，其特征在于，所述步骤 c3 进一步包括：

c31. 获取已分类为不同主题的词义选项的多义词条样本库；

c32. 获取所述词义选项的目录信息；

c33. 根据所述词义选项的目录信息提取所述词义选项的目录特征；

c34. 根据所述词义选项的目录特征进行机器建模，生成包括所述词义选项的相邻的所述目录信息的所述目录特征的关联度关系的机器模型。

14. 根据权利要求 13 所述的词条处理方法，其特征在于，所述词义选项的相邻的所述目录信息的所述目录特征的关联度关系包括相同词汇的数量、相同词汇的词汇属性、相同词汇所占目录信息的比例、关联词汇的数量、关联词汇的词汇属性、关联词汇所占目录信息的比例之一或其组合。

15. 根据权利要求 10 所述的词条处理方法，其特征在于，在所述步骤 c4 中，进行关联度计算的方法包括计算相邻的所述目录信息的所述目录特征中相同词汇的数量。

16. 根据权利要求 15 所述的词条处理方法，其特征在于，在所述步骤 c4 中，进行关联度计算的方法还包括判断所述相同词汇的词汇属性。

17. 根据权利要求 15 所述的词条处理方法，其特征在于，在所述步骤 c4 中，进行关联度计算的方法还包括计算所述相同词汇所占目录信息的比例。

18. 根据权利要求 10 所述的词条处理方法，其特征在于，在所述步骤 c4 中，进行关联度计算的方法包括计算相邻的所述目录信息的所述目录特征中关联词汇的数量。

19. 根据权利要求 18 所述的词条处理方法，其特征在于，在所述步骤 c4 中，进行关联度计算的方法还包括判断所述关联词汇的词汇属性。

20. 根据权利要求 18 所述的词条处理方法，其特征在于，在所述步骤 c4 中，进行关联度计算的方法还包括计算所述关联词汇所占目录信息的比例。

21. 根据权利要求 10 所述的词条处理方法，其特征在于，所述步骤 c5 进一步包括：

c51. 根据关联度计算结果将前后相邻的所述目录信息划分为相关和无关；

c52. 将起始目录以及与在前目录信息相关的所述目录信息标记为第一标记；

c53. 将与在前目录信息无关的所述目录信息标记为第二标记。

22. 一种开放式知识共享平台，其特征在于，所述开放式知识共享平台包括：

词条获取模块，获取词条及词条内容；

多义词判断模块，判断所述词条是否为涉及不同主题的多义词；

词义选项划分模块，若所述词条为涉及不同主题的多义词，则将涉及不同主题的所述词条内容分别划分到对应的词义选项下。

23. 根据权利要求 22 所述的开放式知识共享平台，其特征在于，所述词条内容包括多

个目录,所述多义词判断模块根据所述目录中的目录信息判断所述词条是否为涉及不同主题的多义词。

24. 根据权利要求 22 所述的开放式知识共享平台,其特征在于,所述词条获取模块进一步获取所述词条的分类信息,所述多义词判断模块根据所述分类信息判断所述词条是否为涉及不同主题的多义词。

25. 根据权利要求 22 所述的开放式知识共享平台,其特征在于,所述多义词判断模块包括 :

词条特征提取模块,对所述词条内容进行特征提取,以获取多个词条特征;

标注特征集合获取模块,获取标注特征集合,所述标注特征集合包括多个带有权重值的标注特征;

词条特征权重计算模块,根据所述标注特征分别为每一所述词条特征分配对应的权重值;以及

阈值判断模块,对所述多个词条特征的权重值进行求和,并将权重值求和高于阈值的词条作为多义词。

26. 根据权利要求 25 所述的开放式知识共享平台,其特征在于,所述标注特征集合获取模块进一步包括 :

标注语料获取模块,获取包括多个多义词样本和非多义词样本的标注语料;

标注特征提取模块,从所述标注语料中提取多个所述标注特征;

标注特征权重计算模块,根据所述标注特征在所述多义词样本和所述非多义词样本的出现情况为所述标注特征分别分配对应的权重值。

27. 根据权利要求 26 所述的开放式知识共享平台,其特征在于,所述标注特征权重计算模块计算所述标注特征在所述多义词样本中出现的次数 M,计算所述标注特征在所述非多义词样本中出现的次数 N,并计算所述权重值等于 $M / (M+N)$ 。

28. 根据权利要求 22 所述的开放式知识共享平台,其特征在于,所述词义选项划分模块以所述词条的一级目录以及所述一级目录下的目录内容作为基本预测单元,预测所述基本预测单元是否属于不同主题,并根据预测结果将属于不同主题的所述基本预测单元分别划分到对应的词义选项。

29. 根据权利要求 28 所述的开放式知识共享平台,其特征在于,所述词义选项划分模块根据所述一级目录中的目录信息进行预测。

30. 根据权利要求 22 所述的开放式知识共享平台,其特征在于,所述词义选项划分模块进一步包括 :

目录信息获取模块,获取所述词条的目录信息;

目录特征提取模块,根据所述目录信息提取目录特征;

机器模型获取模块,获取含目录特征关联度关系的机器模型;

关联度计算模块,根据所提取的所述目录特征,应用所述机器模型对前后相邻的所述目录信息的所述目录特征进行关联度计算,以及

标记模块,根据关联度计算结果对所述目录信息进行标记。

31. 根据权利要求 30 所述的开放式知识共享平台,其特征在于,所述机器模型获取模块进一步包括 :

词义选项样本获取模块,获取已分类为不同主题的词义选项的多义词条样本库;

词义选项目录信息获取模块,获取所述词义选项的目录信息;

词义选项目录特征提取模块,根据所述词义选项的目录信息提取所述词义选项的目录特征;

机器建模模块,根据所述词义选项的目录特征进行机器建模,生成包括所述词义选项的相邻的所述目录信息的所述目录特征的关联度关系的机器模型。

32. 根据权利要求 31 所述的开放式知识共享平台,其特征在于,所述词义选项的相邻的所述目录信息的所述目录特征的关联度关系包括相同词汇的数量、相同词汇的词汇属性、相同词汇所占目录信息的比例、关联词汇的数量、关联词汇的词汇属性、关联词汇所占目录信息的比例之一或其组合。

33. 根据权利要求 30 所述的开放式知识共享平台,其特征在于,所述关联度计算模块进行关联度计算的方法包括计算相邻的所述目录信息的所述目录特征中相同词汇的数量。

34. 根据权利要求 33 所述的开放式知识共享平台,其特征在于,所述关联度计算模块进行关联度计算的方法还包括判断所述相同词汇的词汇属性。

35. 根据权利要求 33 所述的开放式知识共享平台,其特征在于,所述关联度计算模块进行关联度计算的方法还包括计算所述相同词汇所占目录信息的比例。

36. 根据权利要求 30 所述的开放式知识共享平台,其特征在于,所述关联度计算模块进行关联度计算的方法包括计算相邻的所述目录信息的所述目录特征中关联词汇的数量。

37. 根据权利要求 36 所述的开放式知识共享平台,其特征在于,所述关联度计算模块进行关联度计算的方法还包括判断所述关联词汇的词汇属性。

38. 根据权利要求 36 所述的开放式知识共享平台,其特征在于,所述关联度计算模块进行关联度计算的方法还包括计算所述关联词汇所占目录信息的比例。

39. 根据权利要求 30 所述的开放式知识共享平台,其特征在于,所述标记模块进一步包括:

关联度分类模块,根据关联度计算结果将前后相邻的所述目录信息划分为相关和无关;

第一标记模块,将起始目录以及与在前目录信息相关的所述目录信息标记为第一标记;

第二标记模块,将与在前目录信息无关的所述目录信息标记为第二标记。

一种开放式知识共享平台及其词条处理方法

【技术领域】

[0001] 本发明涉及网络技术,特别涉及一种开放式知识共享平台及其词条处理方法。

【背景技术】

[0002] 随着互联网技术的发展,开放式知识共享平台服务于互联网络中业已广泛应用。知识共享平台由大量用户对各种词条进行编辑,用以帮助有需要的人进行了解。但开放式知识共享平台经常存在涉及不同主题的多义词,例如词条“孙悦”可能涉及到“歌手孙悦”,也可能涉及到“球员孙悦”。再例如词条“苹果”可能涉及到植物、公司、国内电影以及国外电影等等。目前,现有的开放式知识共享平台是将涉及不同主题的多义词的词条内容统一呈现给用户,用户需要在众多词条内容中寻找自己想要的解释,导致用户的浏览体验不佳。

【发明内容】

[0003] 有鉴于此,本发明提供了一种开放式知识共享平台及其词条处理方法,将涉及不同主题的词条内容划分到不同词义选项下,使得开放式知识共享平台内容的粒度更细,更便于对同名不同主题的内容进行编辑完善和有针对性的引入扩展内容,从而提升用户的浏览体验。

[0004] 本发明提供了一种开放式知识共享平台的词条处理方法,其特征在于,词条处理方法包括 :a. 获取词条及词条内容 ;b. 判断词条是否为涉及不同主题的多义词 ;c. 若词条为涉及不同主题的多义词,则将涉及不同主题的词条内容分别划分到对应的词义选项下。

[0005] 根据本发明之一优选实施方式,词条内容包括多个目录,在步骤 b 中,根据目录中的目录信息判断词条是否为涉及不同主题的多义词。

[0006] 根据本发明之一优选实施方式,在步骤 a 中,进一步获取词条的分类信息,在步骤 b 中,根据分类信息判断词条是否为涉及不同主题的多义词。

[0007] 根据本发明之一优选实施方式,步骤 b 包括 :b1. 对词条内容进行特征提取,以获取多个词条特征 ;b2. 获取标注特征集合,标注特征集合包括多个带有权重值的标注特征 ;b3. 根据标注特征分别为每一词条特征分配对应的权重值 ;以及 b4. 对多个词条特征的权重值进行求和,并将权重值求和高于阈值的词条作为多义词。

[0008] 根据本发明之一优选实施方式,步骤 b2 进一步包括 :b21. 获取包括多个多义词样本和非多义词样本的标注语料 ;b22. 从标注语料中提取多个标注特征 ;b23. 根据标注特征在多义词样本和非多义词样本的出现情况为标注特征分别分配对应的权重值。

[0009] 根据本发明之一优选实施方式,在步骤 b23 中,计算标注特征在多义词样本中出现的次数 M,计算标注特征在非多义词样本中出现的次数 N,并计算权重值等于 $M/(M+N)$ 。

[0010] 根据本发明之一优选实施方式,在步骤 c 中,以词条的一级目录以及一级目录下的目录内容作为基本预测单元,预测基本预测单元是否属于不同主题,并根据预测结果将属于不同主题的基本预测单元分别划分到对应的词义选项。

[0011] 根据本发明之一优选实施方式,在步骤 c 中,根据一级目录中的目录信息进行预

测。

[0012] 根据本发明之一优选实施方式,步骤 c 进一步包括:c1. 从基本预测单元中提取多个目录特征;c2. 根据目录特征判断基本预测单元之间的关联度,并根据关联度产生预测结果。

[0013] 根据本发明之一优选实施方式,步骤 c 进一步包括:c1. 获取词条的目录信息;c2. 根据目录信息提取目录特征;c3. 获取含目录特征关联度关系的机器模型;c4. 根据所提取的目录特征,应用机器模型对前后相邻的目录信息的目录特征进行关联度计算;c5. 根据关联度计算结果对目录信息进行标记。

[0014] 根据本发明之一优选实施方式,步骤 c2 进一步包括:在提取目录特征时,先进行分词。

[0015] 根据本发明之一优选实施方式,分词的方法包括:正向匹配分词、反向匹配分词、正向反向匹配分词、基于全切分词图的分词、最大熵马尔科夫模型分词、最大熵分词或条件随机场分词。

[0016] 根据本发明之一优选实施方式,步骤 c3 进一步包括:c31. 获取已分类为不同主题的词义选项的多义词条样本库;c32. 获取词义选项的目录信息;c33. 根据词义选项的目录信息提取词义选项的目录特征;c34. 根据词义选项的目录特征进行机器建模,生成包括词义选项的相邻的目录信息的目录特征的关联度关系的机器模型。

[0017] 根据本发明之一优选实施方式,词义选项的相邻的目录信息的目录特征的关联度关系包括相同词汇的数量、相同词汇的词汇属性、相同词汇所占目录信息的比例、关联词汇的数量、关联词汇的词汇属性、关联词汇所占目录信息的比例之一或其组合。

[0018] 根据本发明之一优选实施方式,在步骤 c4 中,进行关联度计算的方法包括计算相邻的目录信息的目录特征中相同词汇的数量。

[0019] 根据本发明之一优选实施方式,在步骤 c4 中,进行关联度计算的方法还包括判断相同词汇的词汇属性。

[0020] 根据本发明之一优选实施方式,在步骤 c4 中,进行关联度计算的方法还包括计算相同词汇所占目录信息的比例。

[0021] 根据本发明之一优选实施方式,在步骤 c4 中,进行关联度计算的方法包括计算相邻的目录信息的目录特征中关联词汇的数量。

[0022] 根据本发明之一优选实施方式,在步骤 c4 中,进行关联度计算的方法还包括判断关联词汇的词汇属性。

[0023] 根据本发明之一优选实施方式,在步骤 c4 中,进行关联度计算的方法还包括计算关联词汇所占目录信息的比例。

[0024] 根据本发明之一优选实施方式,步骤 c5 进一步包括:c51. 根据关联度计算结果将前后相邻的目录信息划分为相关和无关;c52. 将起始目录以及与在前目录信息相关的目录信息标记为第一标记;c53. 将与在前目录信息无关的目录信息标记为第二标记。

[0025] 本发明还提供一种开放式知识共享平台,开放式知识共享平台包括:词条获取模块,获取词条及词条内容;多义词判断模块,判断词条是否为涉及不同主题的多义词;词义选项划分模块,若词条为涉及不同主题的多义词,则将涉及不同主题的词条内容分别划分到对应的词义选项下。

[0026] 根据本发明之一优选实施方式，词条内容包括多个目录，多义词判断模块根据目录中的目录信息判断词条是否为涉及不同主题的多义词。

[0027] 根据本发明之一优选实施方式，词条获取模块进一步获取词条的分类信息，多义词判断模块根据分类信息判断词条是否为涉及不同主题的多义词。

[0028] 根据本发明之一优选实施方式，多义词判断模块包括：词条特征提取模块，对词条内容进行特征提取，以获取多个词条特征；标注特征集合获取模块，获取标注特征集合，标注特征集合包括多个带有权重值的标注特征；词条特征权重计算模块，根据标注特征分别为每一词条特征分配对应的权重值；以及阈值判断模块，对多个词条特征的权重值进行求和，并将权重值求和高于阈值的词条作为多义词。

[0029] 根据本发明之一优选实施方式，标注特征集合获取模块进一步包括：标注语料获取模块，获取包括多个多义词样本和非多义词样本的标注语料；标注特征提取模块，从标注语料中提取多个标注特征；标注特征权重计算模块，根据标注特征在多义词样本和非多义词样本的出现情况为标注特征分别分配对应的权重值。

[0030] 根据本发明之一优选实施方式，标注特征权重计算模块计算标注特征在多义词样本中出现的次数 M ，计算标注特征在非多义词样本中出现的次数 N ，并计算权重值等于 $M/(M+N)$ 。

[0031] 根据本发明之一优选实施方式，词义选项划分模块以词条的一级目录以及一级目录下的目录内容作为基本预测单元，预测基本预测单元是否属于不同主题，并根据预测结果将属于不同主题的基本预测单元分别划分到对应的词义选项。

[0032] 根据本发明之一优选实施方式，词义选项划分模块根据一级目录中的目录信息进行预测。

[0033] 根据本发明之一优选实施方式，词义选项划分模块进一步包括：目录信息获取模块，获取词条的目录信息；目录特征提取模块，根据目录信息提取目录特征；机器模型获取模块，获取含目录特征关联度关系的机器模型；关联度计算模块，根据所提取的目录特征，应用机器模型对前后相邻的目录信息的目录特征进行关联度计算，以及标记模块，根据关联度计算结果对目录信息进行标记。

[0034] 根据本发明之一优选实施方式，机器模型获取模块进一步包括：词义选项样本获取模块，获取已分类为不同主题的词义选项的多义词条样本库；词义选项目录信息获取模块，获取词义选项的目录信息；词义选项目录特征提取模块，根据词义选项的目录信息提取词义选项的目录特征；机器建模模块，根据词义选项的目录特征进行机器建模，生成包括词义选项的相邻的目录信息的目录特征的关联度关系的机器模型。

[0035] 根据本发明之一优选实施方式，词义选项的相邻的目录信息的目录特征的关联度关系包括相同词汇的数量、相同词汇的词汇属性、相同词汇所占目录信息的比例、关联词汇的数量、关联词汇的词汇属性、关联词汇所占目录信息的比例之一或其组合。

[0036] 根据本发明之一优选实施方式，关联度计算模块进行关联度计算的方法包括计算相邻的目录信息的目录特征中相同词汇的数量。

[0037] 根据本发明之一优选实施方式，关联度计算模块进行关联度计算的方法还包括判断相同词汇的词汇属性。

[0038] 根据本发明之一优选实施方式，关联度计算模块进行关联度计算的方法还包括计

算相同词汇所占目录信息的比例。

[0039] 根据本发明之一优选实施方式，关联度计算模块进行关联度计算的方法包括计算相邻的目录信息的目录特征中关联词汇的数量。

[0040] 根据本发明之一优选实施方式，关联度计算模块进行关联度计算的方法还包括判断关联词汇的词汇属性。

[0041] 根据本发明之一优选实施方式，关联度计算模块进行关联度计算的方法还包括计算关联词汇所占目录信息的比例。

[0042] 根据本发明之一优选实施方式，标记模块进一步包括：关联度分类模块，根据关联度计算结果将前后相邻的目录信息划分为相关和无关；第一标记模块，将起始目录以及与在前目录信息相关的目录信息标记为第一标记；第二标记模块，将与在前目录信息无关的目录信息标记为第二标记。

[0043] 通过上述所提供的技术方案，本发明提供了一种开放式知识共享平台及其词条处理方法，可将涉及不同主题的词条内容划分到不同词义选项下，使得开放式知识共享平台内容的粒度更细，更便于对同名不同主题的内容进行编辑完善和有针对性的引入扩展内容，从而提升用户的浏览体验。

【附图说明】

[0044] 图 1 是本发明的开放式知识共享平台的多义词展现方法的流程示意图；

[0045] 图 2 是本发明的开放式知识共享平台的示意框图；

[0046] 图 3 是本发明的开放式知识共享平台的词条处理方法的流程示意图；

[0047] 图 4 是本发明的开放式知识共享平台的多义词判定方法的流程示意图

[0048] 图 5 是本发明的开放式知识共享平台的标注特征集合获取方法的流程示意图。

[0049] 图 6 是本发明的开放式知识共享平台的多义词条的词义选项划分方法的流程示意图；

[0050] 图 7 是本发明的开放式知识共享平台的词条处理装置的示意框图；

[0051] 图 8 是本发明的开放式知识共享平台的多义词判定装置的示意框图

[0052] 图 9 是本发明的开放式知识共享平台的标注特征集合获取装置的示意框图。

[0053] 图 10 是本发明的开放式知识共享平台的多义词条的词义选项划分装置的示意框图。

【具体实施方式】

[0054] 为了使本发明的目的、技术方案和优点更加清楚，下面结合附图和具体实施方式对本发明进行详细描述。

[0055] 首先请参见图 1，图 1 是本发明的开放式知识共享平台的多义词展现方法的流程示意图。

[0056] 在步骤 10 中，将同一词条的涉及不同主题的词条内容分别与对应的词义选项进行关联。例如，在本步骤中，将词条“孙悦”的涉及到“歌手孙悦”的词条内容划分并关联到词义选项“歌手孙悦”下，并将涉及到“球员孙悦”的词条内容划分并关联到词义选项“球员孙悦”下。多义词的具体判断过程以及词条内容的具体划分过程将在下文中进行描述。

[0057] 在步骤 11 中,接收用户的查询请求。具体来说,用户通过浏览器发送查询请求,该查询请求经互联网传送到开放式知识共享平台。

[0058] 在步骤 12 中,查询与查询请求相匹配的词条。

[0059] 在步骤 13 中,输出并在浏览器上呈现出与相匹配的词条对应的多个词义选项。具体来说,开放式知识共享平台在接收到查询请求后,在数据库内查询与该查询请求相匹配的词条。若该词条为涉及不同主题的多义词,则输出并在浏览器中呈现与相匹配词条对应的多个词义选项。例如,当用户输入的查询请求为“孙悦”时,开放式知识共享平台输出并在浏览器上呈现出“歌手孙悦”以及“球员孙悦”两个词义选项。

[0060] 在步骤 14 中,根据用户针对词义选项的请求展现相关联的词条内容。具体来说,用户通过词义选项确定自己所感兴趣的主题,进而点击对应的词义选项。开放式知识共享平台则将与该词义选项相关联的词条内容输出到浏览器上,进而展现给用户。例如,用户感兴趣的是“歌手孙悦”,用户点击词义选项“歌手孙悦”,则开放式知识共享平台将与“歌手孙悦”相关的词条内容输出到浏览器上,进而展现给用户。

[0061] 在步骤 13 中,优选根据用户行为对浏览器上呈现的多个词义选项进行排序,使得关注度高的词义选项排在该多个词义选项的前列,由此进一步提高用户的浏览体验。

[0062] 例如,可根据词条内容的展现量、用户对词条内容的浏览时间或词条内容的点击量与展现量之比对多个词义选项进行排序。其中,词条内容的展现量是指根据用户针对词义选项的请求展现相关联的词条内容的次数。一般来说,词条内容的展现量越多证明该词条内容受到的关注度越高,其所对应的词义选项排名应该越靠前。用户在词条内容的浏览时间是指用户在该词条内容展现后到该词条内容关闭之前所花费的时间,用户对词条内容的浏览时间越长,也证明用户对该词条的关注度越高,其所对应的词义选项排名应该越靠前。词条内容的点击量是指用户对所展现的词条内容中的标题、图片或链接等内容的点击次数。词条内容的点击量与展现量之比越高,则也证明用户对该词条的关注度越高,其所对应的词义选项排名应该越靠前。

[0063] 在优选实施方式中,综合考虑上述三个衡量标准,来对词义选项进行排序。也就是,根据词义内容的展现量、用户对词义内容的浏览时间以及词义内容的点击量与展现量之比的加权统计结果对多个词义选项进行排序。具体加权统计算法可根据实际需要进行设计。例如,在统计词义内容的展现量时,将浏览时间较短的展现行为赋予较低的权重,再参与到展现量统计中,由此可降低由于用户误操作所产生的展现量对词义选项排序的影响。

[0064] 如图 2 所示,图 2 是本发明的开放式知识共享平台的示意框图。本发明的开放式知识共享平台包括关联模块 20、输入模块 21、匹配模块 22 以及输出模块 23。

[0065] 关联模块 20 将同一词条的涉及不同主题的词条内容分别与对应的词义选项进行关联。例如,关联模块 20 将词条“孙悦”的涉及到“歌手孙悦”的词条内容划分并关联到词义选项“歌手孙悦”下,并将涉及到“球员孙悦”的词条内容划分并关联到词义选项“球员孙悦”下。

[0066] 输入模块 21 接收用户通过浏览器发送的查询请求。匹配模块 22 根据输入模块 21 接收到查询请求在数据库内查询与该查询请求相匹配的词条。若该词条为涉及不同主题的多义词,则输出模块 23 输入该词条对应的多个词义选项,进而在浏览器中呈现该多个词义选项。例如,当用户输入的查询请求为“孙悦”时,输出模块 23 输出并在浏览器上呈现出“歌

手孙悦”以及“球员孙悦”两个词义选项。

[0067] 输入模块 21 进一步接收用户针对某个词义选项的请求，输出模块 23 进一步根据用户针对词义选项的请求输出并进而展现相关联的词条内容。具体来说，用户通过词义选项确定自己所感兴趣的主题，进而点击对应的词义选项。输出模块 23 则将与该词义选项相关联的词条内容输出到浏览器上，进而展现给用户。例如，用户感兴趣的是“歌手孙悦”，用户点击词义选项“歌手孙悦”，则输出模块 23 则将与“歌手孙悦”相关的词条内容输出到浏览器上，进而展现给用户。

[0068] 输出模块 23 优选根据用户行为对浏览器上呈现的多个词义选项进行排序，使得关注度高的词义选项排在该多个词义选项的前列，由此进一步提高用户的浏览体验。

[0069] 例如，可根据词条内容的展现量、用户对词条内容的浏览时间或词条内容的点击量与展现量之比对多个词义选项进行排序。其中，词条内容的展现量是指根据用户针对词义选项的请求展现相关联的词条内容的次数。一般来说，词条内容的展现量越多证明该词条内容受到的关注度越高，其所对应的词义选项排名应该越靠前。用户在词条内容的浏览时间是指用户在该词条内容展现后到该词条内容关闭之前所花费的时间，用户对词条内容的浏览时间越长，也证明用户对该词条的关注度越高，其所对应的词义选项排名应该越靠前。词条内容的点击量是指用户对所展现的词条内容中的标题、图片或链接等内容的点击次数。词条内容的点击量与展现量之比越高，则也证明用户对该词条的关注度越高，其所对应的词义选项排名应该越靠前。

[0070] 在优选实施方式中，综合考虑上述三个衡量标准，来对词义选项进行排序。也就是，根据词义内容的展现量、用户对词义内容的浏览时间以及词义内容的点击量与展现量之比的加权统计结果对多个词义选项进行排序。具体加权统计算法可根据实际需要进行设计。例如，在统计词义内容的展现量时，将浏览时间较短的展现行为赋予较低的权重，再参与到展现量统计中，由此可降低由于用户误操作所产生的展现量对词义选项排序的影响。

[0071] 如图 3 所示，图 3 是本发明的开放式知识共享平台的词条处理方法的流程示意图。

[0072] 在步骤 30 中，获取词条以及词条内容。在优选实施方式中，该词条及词条内容可以是现有开放式知识共享平台中的以目录形式呈现的词条及词条内容。也就是说，该词条内容包括多个目录以及分别位于各目录下的目录内容。其中，目录可以包括多个一级目录，每个一级目录下还可以进一步包括多个二级目录、三级目录等子目录。

[0073] 在步骤 31 中，判断词条是否为涉及不同主题的多义词。判断词条是否为多义词的方法有多种，下面将以几个具体实施方式进行描述。

[0074] 在一实施方式中，根据目录中的目录信息判断词条是否为涉及不同主题的多义词。具体来说，判断不同的目录信息中是否存在涉及不同主题的关键词。例如，如果词条“孙悦”的两个目录信息中分别出现了“歌手孙悦”和“球员孙悦”，由于“歌手”和“球员”涉及到不同的主题，则认为该词条“孙悦”为多义词。再例如，在词条“射雕英雄传”的两个目录信息中分别出现了“1983 版”和“2008 版”，则同样认为词条“射雕英雄传”为多义词。

[0075] 在一实施方式中，在步骤 30 中，进一步获取词条的分类信息，在步骤 31 中，根据词条的分类信息判断该词条是否为不同主题的多义词。例如，词条“苹果”的分类信息中包括“植物”、“电影”以及“公司”三个不同的分类，则认为词条“苹果”为多义词。

[0076] 在另一实施方式中，可根据词条内容通过机器挖掘方法来自动识别词条是否是多

义词。如图 4 所示,图 4 是本发明的开放式知识共享平台的多义词判定方法的流程示意图。

[0077] 在步骤 40 中,对待判定词条的词条内容进行特征提取,以获取多个词条特征。具体来说,对词条内容进行分词及过滤,并将分词及过滤后获得的字词作为词条特征。其中,分词的作用是将词条内容中的汉字序列切分成有意义的字词,以便后续处理。具体分词的方法包括:正向匹配分词、反向匹配分词、正向反向匹配分词、基于全切分词图的分词,最大熵马尔科夫模型分词、最大熵分词或条件随机场分词等,上述分词方法均为本领域公知技术,在此不再赘述。过滤的作用是去除标点符号、助词等无用信息。

[0078] 在步骤 41 中,获取标注特征集合。标注特征集合包括多个带有权重值的标注特征。如图 5 所示,图 5 是本发明的开放式知识共享平台的标注特征集合获取方法的流程示意图。

[0079] 在步骤 50 中,获取包括多个多义词样本和非多义词样本的标注语料。在标注语料中,多义词样本是指已经判定为多义词的词条及词条内容,非多义词样本是指已经判定为非多义词的词条及词条内容。

[0080] 在步骤 51 中,从标注语料中提取多个标注特征。具体来说,分别对每个多义词样本以及每个非多义词样本进行分词及过滤,并将分词及过滤后获得的字词作为标注特征。

[0081] 在步骤 52 中,根据标注特征在多义词样本和非多义词样本的出现情况为标注特征分别分配对应的权重值。具体来说,计算标注特征在多义词样本中出现的次数 M,计算标注特征在非多义词样本中出现的次数 N,并计算该标注特征的权重值等于 $M / (M+N)$ 。通过上述方法可知,如果某个标注特征在多义词样本中出现的次数较多,同时在非多义词样本中出现的次数较少时,该标注特征的权重值就相对较高。如果某个标注特征在多义词样本和非多义词中出现的次数相仿或者在多义词样本中出现的次数较少,同时在非多义词样本中出现的次数较多时,该标注特征的权重值就相对较低。

[0082] 在步骤 42 中,根据标注特征分别为从待判定词条的词条内容提取的每一词条特征分配对应的权重值。具体来说,在标注特征集合中判断是否存在与每一词条特征相同的标注特征,若存在,则将该标注特征的权重值分配给该词条特征。

[0083] 在步骤 43 中,对从待判定词条的词条内容提取的多个词条特征的权重值进行求和,并将权重值求和高于阈值的词条作为多义词。具体来说,如果待判定词条的多个词条特征的权重值越高,则代表该词条特征在多义词样本中出现的次数越高,该待判定词条是多义词的概率就越高。在本实施方式中,具体阈值可根据实际情况进行设置。

[0084] 在步骤 32 中,若判定词条为涉及不同主题的多义词,则将涉及不同主题的词条内容分别划分到对应词义选项下。在一优选实施方式中,以词条的一级目录以及一级目录下的目录内容作为一个基本预测单元,预测基本预测单元是否属于不同主题,并根据预测结果将属于同一主题的基本预测单元划分到同一词义选项。预测基本预测单元是否属于不同主题有多种,下面将以几个具体实施方式进行描述。

[0085] 在一实施方式中,根据一级目录中的目录信息进行预测。例如,如果词条“孙悦”的两个一级目录的目录信息中分别出现了“歌手孙悦”和“球员孙悦”,由于“歌手”和“球员”涉及到不同的主题,则将目录信息中包含“歌手孙悦”的一级目录及其目录内容划分并关联到词义选项“歌手孙悦”下,并将目录信息中包含“球员孙悦”的一级目录及其目录内容划分并关联到词义选项“球员孙悦”下。此外,还可以根据目录信息所体现的用户编辑行为来

进行预测。例如,如果不同的一级目录的目录信息中的第一个字为数字,并连续排列,则将带有数字的一级目录及其目录内容以及下方不带有数字的第一目录及其目录内容分别划分到不同的词义选项下。

[0086] 在一实施方式中,当判定词条为涉及不同主题的多义词时,可以通过机器挖掘的方式将涉及不同主题的词条内容分别划分到对应的词义选项下。如图 6 所示,图 6 是本发明的开放式知识共享平台的多义词条的词义选项划分方法的流程示意图。

[0087] 由于词条中目录通常是按照前后顺序排列的,也就是在通常情况下,多义词条中的相同主题的目录是按照前后顺序排列的,较少出现乱序排列的情况,因此在这种情况下,只需要判断前后目录之间是否是相关的,即可以得知不同主题的目录的分割位置。

[0088] 在步骤 61 中,获取未进行词义选项分类的多义词词条数据。该进行词义选项分类的多义词词条数据可以通过上述图 3 所示的步骤 31 或者图 4 所示多义词判定方法来获得。

[0089] 在步骤 62 中,根据词条中目录的位置获取词条的目录信息。在优选的实施方式中,本发明根据多义词词条中一级目录的位置对词条进行分割,获取每个一级目录的目录信息,目录信息包括一级目录名称以及一级目录下的目录内容等。

[0090] 在步骤 63 中,根据获取的目录信息,从中提取出多个特征。在对目录信息进行特征提取的时候,需要先对词条内容进行分词及过滤,并将分词及过滤后获得的字词作为词条特征。其中,分词的作用是将词条内容中的汉字序列切分成有意义的字词,以便后续处理。具体分词的方法包括:正向匹配分词、反向匹配分词、正向反向匹配分词、基于全切分词图的分词,最大熵马尔科夫模型分词、最大熵分词或条件随机场分词等,上述分词方法均为本领域公知技术,在此不再赘述。过滤的作用是去除标点符号、助词等无用信息。在优选的实施方式中,本发明采用正向最大匹配分词和反向最大匹配分词相结合的方式对分词结果进行校正,以获取正确率较高的分词结果。

[0091] 在步骤 64 中,获取含目录特征关联度关系的机器模型。如图 6 所示,步骤 64 进一步包括:

[0092] 步骤 641,获取已分类为不同主题的词义选项的多义词词条样本库。由于样本是已经分类为不同主题的词义选项的多义词词条,因此每一个词义选项下面的目录所对应的主题是相同的。

[0093] 步骤 642,获取词义选项的目录信息。即获取词义选项下面具有相同主题特性的目录信息。优选的实施方式中,获取词义选项的一级目录信息。

[0094] 步骤 643,根据词义选项的目录信息提取词义选项的目录特征。对具有相同主题特性的目录信息提取对应的目录特征。

[0095] 步骤 644,根据词义选项的目录特征进行机器建模,生成包括词义选项的相邻的目录信息的目录特征的关联度关系的机器模型。由于多义词词条样本库的每一个词条都包括了多个词义选项,通过对同一词义选项下的具有相同主题特性的目录特征进行学习训练,以及对不同词义选项下的具有不同主题特性的目录特征进行学习训练,可以建立包括词义选项的相邻的目录信息的目录特征的关联度关系的机器模型。在优选的实施方式中,词义选项的相邻的目录信息的目录特征的关联度关系包括相同词汇的数量、相同词汇的词汇属性、相同词汇所占目录信息的比例、关联词汇的数量、关联词汇的词汇属性、关联词汇所占目录信息的比例之一或其组合。

[0096] 在步骤 65 中,根据所提取的目录特征,应用机器模型对前后相邻的目录信息的目录特征进行关联度计算。其中,关联度计算可以采用多种方法,并且各种方法之间也可以单独实施或相互结合实施,本发明现举例两种可以单独实施或相互结合实施的方法,但并非用于限制本发明的实施方式。

[0097] 在本发明的一种实施方式中,应用机器模型对前后相邻的目录信息的目录特征中的相同词汇的参数进行计算,通过计算相同词汇的数量、相同词汇所占目录信息的比例,或者对相同词汇的词汇属性进行判断,实现对前后相邻的目录信息的关联度计算。例如,对于文艺作品,特别是影视连续剧、连载小说等,其目录名称相同而内容不同,目录内容中相同词汇出现数量很多,而且词汇属性为名词、动名词等,因此可以据此计算前后目录的关联度。

[0098] 在本发明的另一种实施方式中,应用机器模型对前后相邻的目录信息的目录特征中的关联词汇的参数进行计算,通过计算关联词汇的数量、关联词汇所占目录信息的比例,或者对关联词汇的词汇属性进行判断,实现对前后相邻的目录信息的关联度计算。例如“刘德华”和“朱丽倩”的相关度很高,“刘德华”和“老人”的相关度就低,“歌星”与“专辑”的相关度很高,“歌星”与“战争”的相关度就低。这种词汇相关度可以通过相关词词典或者机器样本学习的方式进行判断。

[0099] 在步骤 66 中,根据关联度计算结果对目录信息进行标记。标记方法包括多种实施方式。在本发明的一种实施方式中,可以根据关联度计算结果将目录信息根据主题进行分类。在本发明的另一种实施方式中,根据关联度计算结果将前后相邻的目录信息划分为相关和无关,将起始目录以及与在前目录信息相关的目录信息标记为第一标记,将与在前目录信息无关的目录信息标记为第二标记。举例来说,一个多义词条包括 6 个目录。通过各个目录和对应的目录内容识别出该目录是不是一个相同主题的词义选项的开始部分,如果是,标记成“B”,如果不是,标记成“I”。这样,6 个目录可能被标记成如“BIBIIB”这样的结果,那么目录 1-2 就是一个词义选项,目录 3-5 是一个词义选项,目录 6 是一个词义选项。这样就实现了对多义词条内具有相同主题的目录的分类。

[0100] 如图 7 所示,图 7 是本发明的开放式知识共享平台的词条处理装置的示意框图。在本实施方式中,词条处理装置包括词条获取模块 70、多义词判断模块 71 以及词义选项划分模块 72

[0101] 词条获取模块 70 用于获取词条以及词条内容。在优选实施方式中,该词条及词条内容可以是现有开放式知识共享平台中的以目录形式呈现的词条及词条内容。也就是说,该词条内容包括多个目录以及分别位于各目录下的目录内容。其中,目录可以包括多个一级目录,每个一级目录下还可以进一步包括多个二级目录、三级目录等子目录。

[0102] 多义词判断模块 71 用于判断词条是否为涉及不同主题的多义词。判断词条是否为多义词的方法有多种,下面将以几个具体实施方式进行描述。

[0103] 在一实施方式中,多义词判断模块 71 根据目录中的目录信息判断词条是否为涉及不同主题的多义词。具体来说,多义词判断模块 71 判断不同的目录信息中是否存在涉及不同主题的关键词。例如,如果词条“孙悦”的两个目录信息中分别出现了“歌手孙悦”和“球员孙悦”,由于“歌手”和“球员”涉及到不同的主题,则认为该词条“孙悦”为多义词。再例如,在词条“射雕英雄传”的两个目录信息中分别出现了“1983 版”和“2008 版”,则同样

认为词条“射雕英雄传”为多义词。

[0104] 在一实施方式中，词条获取模块 70 进一步获取词条的分类信息，多义词判断模块 71 根据词条的分类信息判断该词条是否为不同主题的多义词。例如，词条“苹果”的分类信息中包括“植物”、“电影”以及“公司”三个不同的分类，则认为词条“苹果”为多义词。

[0105] 在另一实施方式中，可根据词条内容通过机器挖掘方法来自动识别词条是否是多义词。如图 8 所示，图 8 是本发明的开放式知识共享平台的多义词判断模块的示意框图。在本实施方式中，多义词判断模块包括词条特征提取模块 80、标注特征集合获取模块 81、词条特征权重计算模块 82 以及阈值判断模块 83

[0106] 词条特征提取模块 80 用于对待判定词条的词条内容进行特征提取，以获取多个词条特征。具体来说，词条特征提取模块 80 对词条内容进行分词及过滤，并将分词及过滤后获得的字词作为词条特征。其中，分词的作用是将词条内容中的汉字序列切分成有意义的字词，以便后续处理。具体分词的方法包括：正向匹配分词、反向匹配分词、正向反向匹配分词、基于全切分词图的分词，最大熵马尔科夫模型分词、最大熵分词或条件随机场分词等，上述分词方法均为本领域公知技术，在此不再赘述。过滤的作用是去除标点符号、助词等无用信息。

[0107] 标注特征集合获取模块 81 用于获取标注特征集合。标注特征集合包括多个带有权重值的标注特征。如图 9 所示，图 9 是本发明的开放式知识共享平台的标注特征集合获取模块的示意框图。在本实施方式中，标注特征集合获取模块包括标注语料获取模块 90、标注特征提取模块 91 以及标注特征权重计算模块 92。

[0108] 标注语料获取模块 90 用于获取包括多个多义词样本和非多义词样本的标注语料。在标注语料中，多义词样本是指已经判定为多义词的词条及词条内容，非多义词样本是指已经判定为非多义词的词条及词条内容。

[0109] 标注特征提取模块 91 用于从标注语料中提取多个标注特征。具体来说，标注特征提取模块 91 分别对每个多义词样本以及每个非多义词样本进行分词及过滤，并将分词及过滤后获得的字词作为标注特征。

[0110] 标注特征权重计算模块 92 用于根据标注特征在多义词样本和非多义词样本的出现情况为标注特征分别分配对应的权重值。具体来说，标注特征权重计算模块 92 计算标注特征在多义词样本中出现的次数 M，计算标注特征在非多义词样本中出现的次数 N，并计算该标注特征的权重值等于 $M / (M+N)$ 。通过上述方法可知，如果某个标注特征在多义词样本中出现的次数较多，同时在非多义词样本中出现的次数较少时，该标注特征的权重值就相对较高。如果某个标注特征在多义词样本和非多义词中出现的次数相仿或者在多义词样本中出现的次数较少，同时在非多义词样本中出现的次数较多时，该标注特征的权重值就相对较低。

[0111] 词条特征权重计算模块 82 用于根据标注特征分别为待判定词条的词条内容提取的每一词条特征分配对应的权重值。具体来说，词条特征权重计算模块 82 在标注特征集合中判断是否存在与每一词条特征相同的标注特征，若存在，则将该标注特征的权重值分配给该词条特征。

[0112] 阈值判断模块 83 用于对待判定词条的词条内容提取的多个词条特征的权重值进行求和，并将权重值求和高于阈值的词条作为多义词。具体来说，如果待判定词条的多个

词条特征的权重值越高，则代表该词条特征在多义词样本中出现的次数越高，该待判定词条是多义词的概率就越高。在本实施方式中，具体阈值可根据实际情况进行设置。

[0113] 若多义词判断模块 71 判定词条为涉及不同主题的多义词，则词义选项划分模块 72 将涉及不同主题的词条内容分别划分到对应词义选项下。在一优选实施方式中，词义选项划分模块 72 以词条的一级目录以及一级目录下的目录内容作为一个基本预测单元，预测基本预测单元是否属于不同主题，并根据预测结果将属于同一主题的基本预测单元划分到同一词义选项。预测基本预测单元是否属于不同主题有多种，下面将以几个具体实施方式进行描述。

[0114] 如图 10 所示，图 10 是本发明的开放式知识共享平台的多义词条的词义选项划分装置的示意框图。词义选项划分模块进一步包括：词条数据获取模块 101、目录信息获取模块 102、目录特征提取模块 103、机器模型获取模块 104、目录相关度计算模块 105 以及标记模块 106。

[0115] 词条数据获取模块 101 用于获取未进行词义选项分类的多义词词条数据。该进行词义选项分类的多义词词条数据可以通过上述图 3 所示的步骤 31 或者图 4 所示多义词判定方法来获得。

[0116] 目录信息获取模块 102 用于根据词条中目录的位置获取词条的目录信息。在优选的实施方式中，本发明根据多义词词条中一级目录的位置对词条进行分割，获取每个一级目录的目录信息，目录信息包括一级目录名称以及一级目录下的目录内容等。

[0117] 目录特征提取模块 103 用于根据获取的目录信息，从中提取出多个特征。在对目录信息进行特征提取的时候，需要先对词条内容进行分词及过滤，并将分词及过滤后获得的字词作为词条特征。在优选的实施方式中，本发明采用正向最大匹配分词和反向最大匹配分词相结合的方式对分词结果进行校正，以获取正确率较高的分词结果。

[0118] 机器模型获取模块 104 用于获取含目录特征关联度关系的机器模型。如图 10 所示，机器模型获取模块 104 进一步包括：样本获取模块 1041、词义选项目录信息获取模块 1042、目录特征提取模块 1043 以及机器建模模块 1044。样本获取模块 1041 用于获取已分类为不同主题的词义选项的多义词条样本库。由于样本是已经分类为不同主题的词义选项的多义词条，因此每一个词义选项下面的目录所对应的主题是相同的。词义选项目录信息获取模块 1042 用于获取词义选项的目录信息。即获取词义选项下面具有相同主题特性的目录信息。优选的实施方式中，获取词义选项的一级目录信息。目录特征提取模块 1043 用于根据词义选项的目录信息提取词义选项的目录特征。对具有相同主题特性的目录信息提取对应的目录特征。机器建模模块 1044 用于根据词义选项的目录特征进行机器建模，生成包括词义选项的相邻的目录信息的目录特征的关联度关系的机器模型。在优选的实施方式中，词义选项的相邻的目录信息的目录特征的关联度关系包括相同词汇的数量、相同词汇的词汇属性、相同词汇所占目录信息的比例、关联词汇的数量、关联词汇的词汇属性、关联词汇所占目录信息的比例之一或其组合。

[0119] 目录相关度计算模块 105 用于根据所提取的目录特征，应用机器模型对前后相邻的目录信息的目录特征进行关联度计算。其中，目录相关度计算模块 105 可以采用多种计算结构实现，并且各种计算结构之间也可以单独实施或相互结合实施。在本发明的一种实施方式中，目录相关度计算模块 105 应用机器模型对前后相邻的目录信息的目录特征中的

相同词汇的参数进行计算,通过计算相同词汇的数量、相同词汇所占目录信息的比例,或者对相同词汇的词汇属性进行判断,实现对前后相邻的目录信息的关联度计算。在本发明的另一种实施方式中,目录相关度计算模块 105 应用机器模型对前后相邻的目录信息的目录特征中的关联词汇的参数进行计算,通过计算关联词汇的数量、关联词汇所占目录信息的比例,或者对关联词汇的词汇属性进行判断,实现对前后相邻的目录信息的关联度计算。

[0120] 标记模块 106 用于根据关联度计算结果对目录信息进行标记。标记方法包括多种实施方式。在本发明的一种实施方式中,可以根据关联度计算结果将目录信息根据主题进行分类。在本发明的另一种实施方式中,根据关联度计算结果将前后相邻的目录信息划分为相关和无关,将起始目录以及与在前目录信息相关的目录信息标记为第一标记,将与在前目录信息无关的目录信息标记为第二标记。标记模块 106 标记模块进一步包括:关联度分类模块、第一标记模块以及第二标记模块。关联度分类模块根据关联度计算结果将前后相邻的目录信息划分为相关和无关。第一标记模块将起始目录以及与在前目录信息相关的目录信息标记为第一标记。第二标记模块将与在前目录信息无关的目录信息标记为第二标记。

[0121] 通过上述所提供的技术方案,本发明提供了一种开放式知识共享平台及其多义词展现方法,可将多义词中不同主题的词义选项显示出来,由用户进行选择,提高用户体验。

[0122] 以上所述仅为本发明的较佳实施方式而已,并不用以限制本发明,凡在本发明的精神和原则之内,所做的任何修改、等同替换、改进等,均应包含在本发明保护的范围之内。

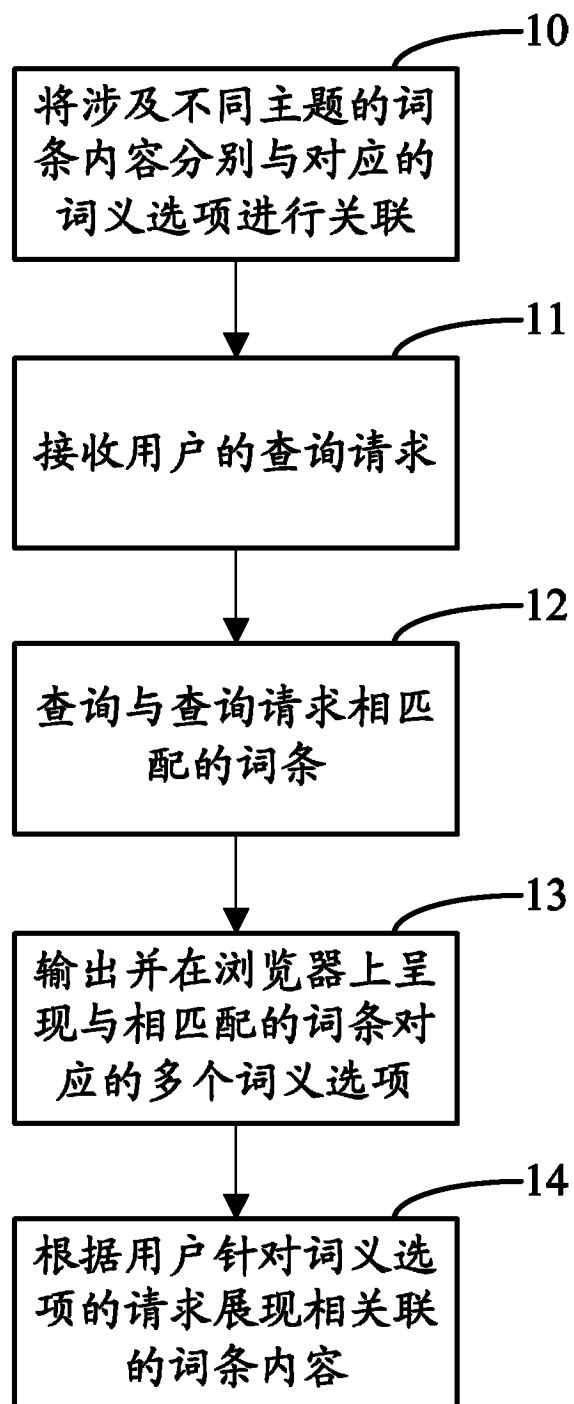


图 1

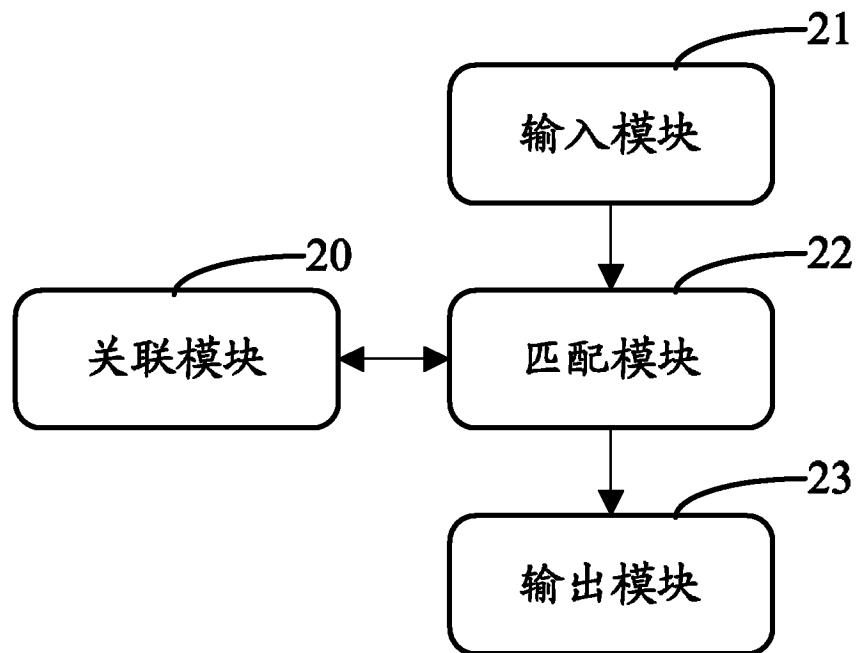


图 2

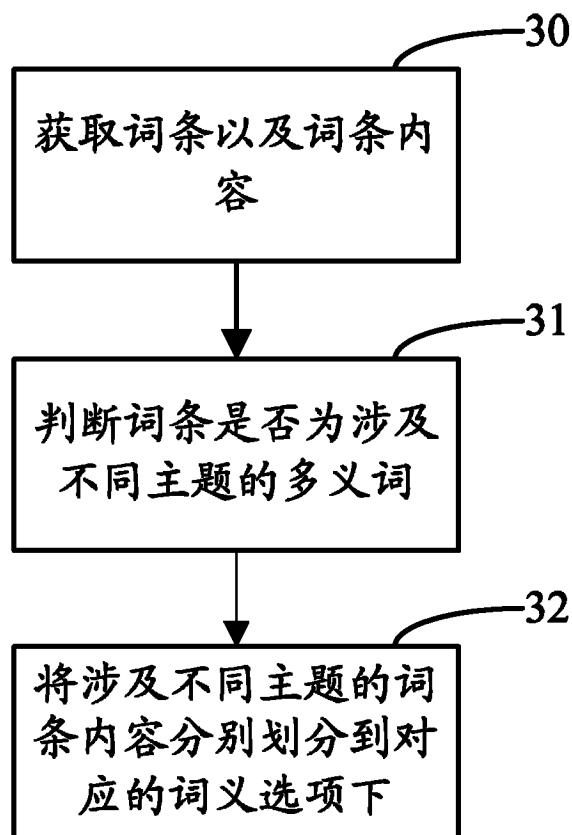
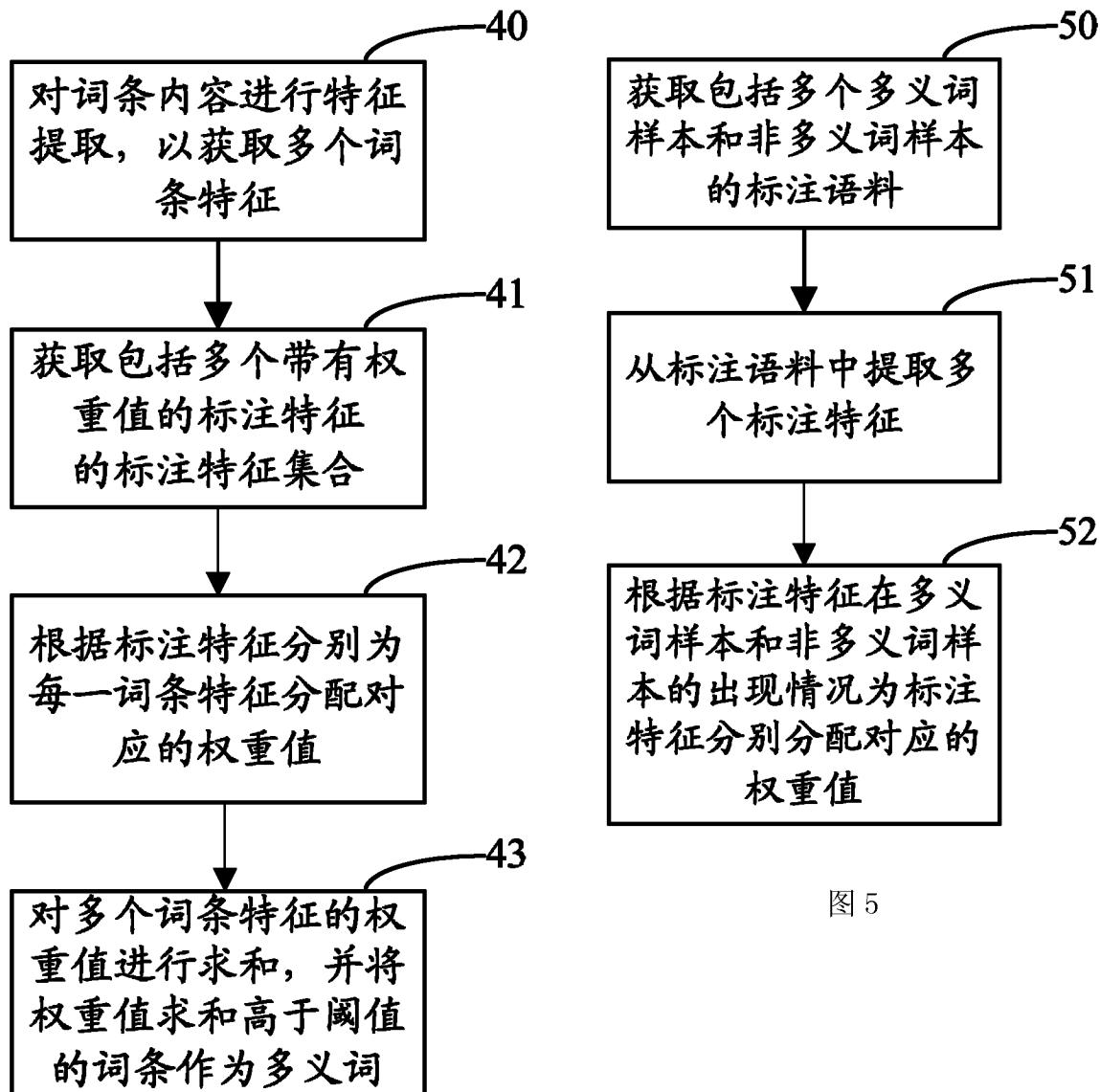


图 3



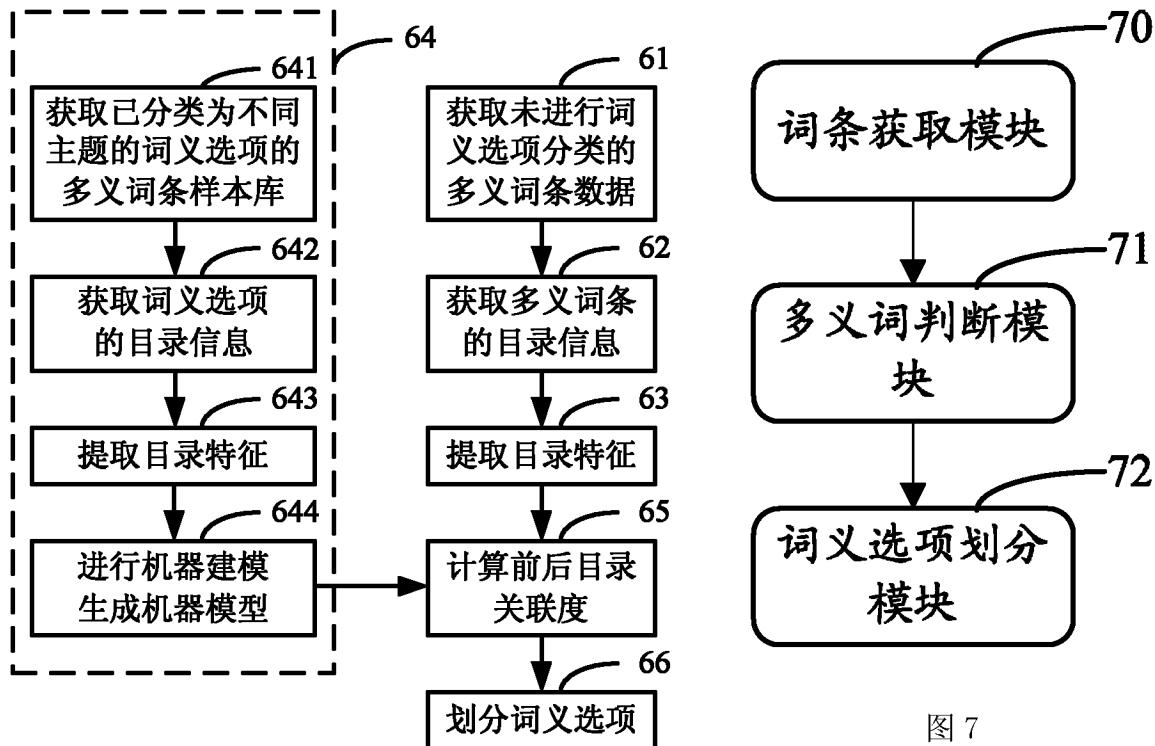


图 7

图 6

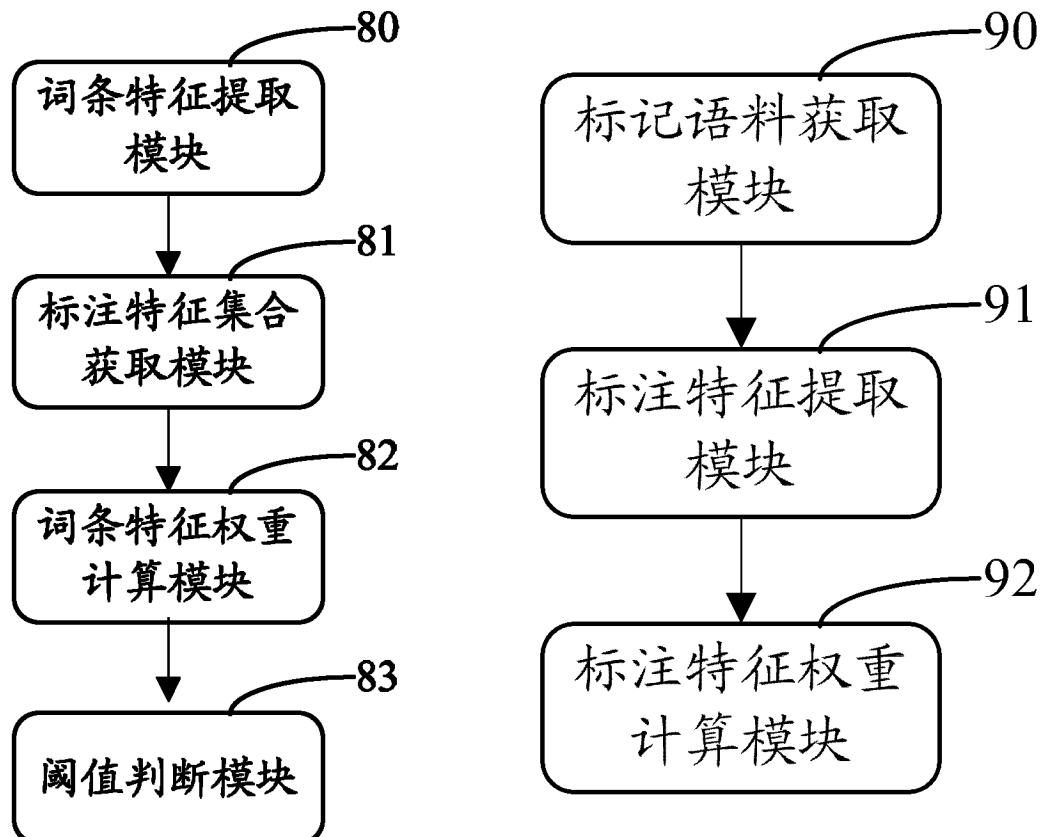


图 9

图 8

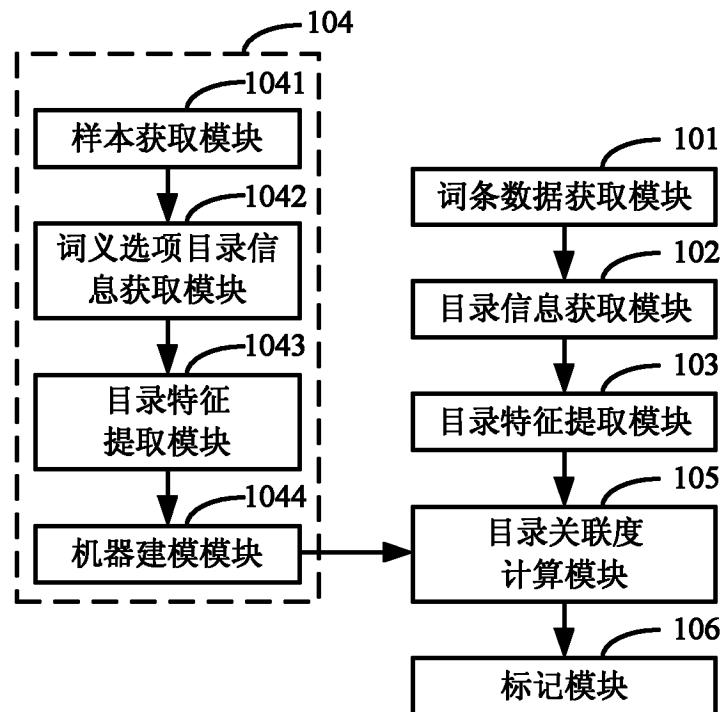


图 10