第36卷 第20期 计 算 机 工 程 2010年10月
Vol.36 No.20 Computer Engineering October 2010
·软件技术与数据库· 文章编号：1000—3428(2010)20—0043—03 文献标识码：A 中图分类号：TP18

文本机会发现研究综述

孙晓华[1,2]，刘大昕[1]，张健沛[1]，徐悦竹[1]
(1. 哈尔滨工程大学计算机科学与技术学院，哈尔滨 150001；2. 哈尔滨理工大学计算机科学与技术学院，哈尔滨 150080)

摘 要：从系统定义、应用环境以及目的等多个角度比较文本机会发现与文本挖掘的不同，分析文本机会发现的研究背景及现状，介绍机会的定义、文本机会发现关键算法、Scenario map分析等文本机会发现的主要研究热点。在总结当前研究的不足的基础上，指出未来文本机会发现的研究方向。

关键词：文本机会发现；机会；文本挖掘；Scenario map分析

Survey of Text Chance Discovery Research

SUN Xiao-hua[1,2], LIU Da-xin[1], ZHANG Jian-pei[1], XU Yue-zhu[1]
(1. School of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China;
2. College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)

【Abstract】 This paper multi-analyses the difference between text chance discovery and text mining from the system definition, application environment and goal etc.. It analyses the background and status, introduces the chance definition, text chance discovery key algorithm and Scenario map analysis etc. main research fields. It indicates the study orientation in the future based on the summarize of current research's limitation.

【Key words】 text chance discovery; chance; text mining; Scenario map analysis

1 概述

Ohsawa Y提出了机会发现(chance discovery)的概念和研究方向[1-2]，目的是在动态不可预测的环境中，发现对主体决策具有重要影响的事件或状态。对于那些有利于主体的事件或状态，应该加以利用；反之，则应加以避免。

因为机会发现涉及了很多领域长期以来遇到的瓶颈性问题，所以一经提出就引起了来自不同领域的研究者的关注，并广泛应用于众多的领域，例如大规模地震的征兆信息[3-4]、经济预测[5]等。至今为止已经在国际上召开了多次重要的会议，取得了一系列的重要成果。例如，2001年第1届机会发现国际专题学术讨论会在日本松江召开；2002年第2届机会发现国际专题学术讨论会在日本东京召开，我国学者参加了国际CD&CM学术会议并报告了自己的研究成果；KES'00~KES'07会议等。

2 文本机会发现与文本挖掘

随着信息技术的迅速发展，大容量存储媒介的进步导致了有效信息数量的爆炸。文本是一种重要的信息承载方式。这些信息以多种形式出现，如电子邮件、新闻等。为此，人们需要在大量的、高维的、无结构数据中进行分析，以得到可以实际利用的数据。

2.1 文本机会发现

Ohsawa Y认为机会发现是发现机会(chance)的过程，不是被机会发现。机会是对决策具有重要影响作用的事件或状态，这样的事件或状态可以被认为是机遇或危机。对于机遇应积极加以促进，而对于被发现的危机则应采取措施加以阻止。机会发现就是对机会的获取并对其含义进行解释的过程。因此，机会的发现是机会发现研究的重点。

当决策者将自己的思想以文档记录下来时，文档的核心内容就与他头脑中的真实想法直接对应。获取文档中表示一个人思想或行为的信息就可帮助了解作者的真实意图。文档中用于表示主要论点的词通常称之为关键词。关键词是概述一个文档的简明方法，并且给出了一个文档内容的更高层次的总结。关键词对大量文本挖掘应用提供了丰富的语义信息。例如主题搜索、文档分析。对于一篇文档，好的关键词能为用户理解文档含义提供许多方便。因此，如何高效地在大量数据中找出表示作者意图的关键词是进行文档处理的关键问题。

为解决此问题，目前有多种文档自动检索方法，例如统计索引、利用自然语言分析进行索引等。这些方法都是将文档中频繁出现的有意义的词作为关键词。但实质上除了这些高频词外，还存在一些表达了作者主要观点的低频词，这些词在分析作者意图时起着至关重要的作用，而传统方法对这些低频词的查找却是无效的。在文本机会发现中，将机会定义为低频出现的关键词。文本机会发现的目的就是要在大量的数据集中找出那些表达了作者的主要观点，但是出现频率却并不高的关键词，即机会。

2.2 文本机会发现与文本挖掘的区别

文本挖掘(text mining)是对具有丰富语义的文本进行分

基金项目：国家自然科学基金资助项目(60873037)；黑龙江省教育厅科学技术研究基金资助项目(11531049)
作者简介：孙晓华(1974－)，女，博士，主研方向：数据库系统及应用，数据挖掘，人工智能；刘大昕、张健沛，教授；徐悦竹，博士
收稿日期：2010-04-10 **E-mail：**corliss@hrbust.edu.cn

析从而发现隐含的，有潜在使用价值知识的过程，是数据挖掘的一个新兴主题。

设 $Y = f(X) = f(x_1, x_2, \cdots, x_n)$ 定义了一个动态确定性系统，该系统对每个输入 X 产生惟一的输出 Y。在这样一个系统中进行数据挖掘通常是通过数据分析获得系统 f 的一个客观描述。这种客观描述的近似有效性是由大量样本数据集支持的。对于 f，机会发现的目的是为了获得系统 f 已经获得的输入以及系统 f 的某个局部状态。这些输入/局部状态是不可预期的稀少事件出现的新特征，这些特征可能对系统 f 的未来状态产生重要影响，或其他未被意识到的局部状态会因这些输入/局部状态而发生显著的变化[6]。

文本机会发现与文本挖掘的区别体现如下：

(1)文本机会发现侧重于确定稀少的但是具有重要意义的关键词。而依据文本挖掘的概念，找到的机会可能被看作是额外事物或噪音数据。

(2)机会发现是一个强调人机交互作用的系统。人的经验、焦点对机会的发现起着至关重要的作用。而数据挖掘工具却强调其独立、高智能地找到潜在的变化规律。

(3)机会发现强调是在动态不可预测的环境中，找出对主体将来决策具有重大影响作用的潜在的、隐含的因果关系。而数据挖掘侧重于在静态环境下找到数据中潜在的变化规律，这个规律表明了主体在未来可能的变化趋势。

但同时，由于人类和他们所处的环境是无法确定的、存在不引人注意的因素的复杂环境，因此机会发现又离不开数据挖掘的支持。数据挖掘提供了在数据集中找出对未来具有重要影响作用的、不被注意的事件的方法。

3 文本机会发现的研究现状

机会发现技术从提出开始，就不断地应用于文档数据分析领域。目前，面向文本的机会发现的研究主要集中于机会的提取算法及可视化技术、机会的定义、Scenario 分析及机会发现的应用等方面。

3.1 机会的定义

3.1.1 Ohsawa Y 的定义

对于什么是“机会”，基于 chance 的基本含义及其机会发现中的作用，Ohsawa Y 对 Chance 的概念作了初步的描述[1]：Chance 是对人的决策制定过程具有重要影响作用的事件或状态，这样的事件或状态可以被认为是 Opportunity 或 Risk。

3.1.2 基于溯因推理的定义

Abe 从溯因推理的角度来研究机会发现。他认为一个事件或状态被认为是机会，则该事件或状态属于下面 2 种情形之一：(1)机会是一些未知假设集合，导致一个原本可以被解释的现象现在不能被解释，这些缺失的假设就可以被认为是机会。(2)机会本身是一些已知的事实，但是对于怎样使用它们去解释现象是未知的，即一组规则缺失。在这种情况下，规则可以通过溯因推理的方法生成。通过溯因推理生成的规则被认为是机会[7]。显然此定义比较符合机会的直观定义，但是此定义无法在计算机内有效地表示。

3.1.3 基于 Lm4c 的定义

中国科技大学陈小平教授及其课题组利用 Lm4c 系统给出了候选机会[2]的定义：一个事件/状态 Ψ 是关于 Agent 目标 φ 的候选 Chance，当且仅当 $\Psi \in \{\varphi\}$ 或 $\neg\Psi \in \{\varphi\}$。

3.1.4 基于遗传算法的机会定义

文献[8]基于遗传算法给出了文档中词的定义。并将其应用于算法 KeyGraph 之中，验证了 KeyGraph 算法可适用于遗传算法，实现了关键词的提取。

机会首先是一个事件或状态，而事件/状态如何定义或描述，在上述方法中都没有给出确切的定义。同时，这些定义方法都是基于特定的理论或环境而制定，因此缺乏普遍适用性。

3.2 文本机会发现算法

3.2.1 KeyGraph 算法

文本机会发现算法中最著名的系统是 KeyGraph[9-10]。KeyGraph 是由 Ohsawa Y 提出的对文档进行索引概括的一种可视化工具，可以用来摘取文档中隐含的必要事件，并从中找出其因果结构，即 Chance。该算法基于图的分割形成与作者意图相一致的群。作者的意图利用关键词来表达，关键词是根据词与群间的关系的统计计算来进行选择。

3.2.2 Data Crystallization 算法

KeyGraph 是基于给定数据集的，用于找出数据集中潜在的或隐含的因果关系。但是，如果该数据集中不能包含全部数据，就会使得计算结果出现偏差，导致最终提取的关键项不准确。Data Crystallization[11]是用来解决数据集数据不完整的情况下，准确获取机会的方法。该方法首先通过 KeyGraph 算法对现有数据进行计算，形成比较粗糙的 KeyGraph 图，然后将与未观测数据相一致的“哑元”作为节点加入到 KeyGraph 图中，再经 KeyGraph 进行可视化，建立“哑元”与图中现有结点间的联系。

3.2.3 KeyWorld 算法

KeyWorld 算法[12]以小世界模型为基础对学术论文形式的文本信息进行可视化组织，以小世界模型的特征路径长度以及聚类系数为基础的关键字自动抽取算法。

上述算法都是以文档中高频词为基础，通过一定的选择标准获取与这些高频词联系紧密的低频词(即候选机会)而实现。它们都较好地实现了一篇文档中高效准确地获取关键词(机会)。但是，一旦算法面向的文档数量增加，其执行效率就会急剧下降，因此，如何提高大规模文本数据集的算法执行效率是急需解决的问题。

3.3 Scenario 分析

通常用来标识机会事件的方法是利用表示人未来的兴趣点的可能的参数值，并使参数能够依据假设的因果过程进行操作而修改，这种方法能够产生一个可能的未来世界状态或这些状态的变化轨迹，称之为 Scenario。对 2 个或多个 Scenario 的比较称为 Scenario 分析[13]。

在文本机会发现中，本文依据发现算法得到了相关数据集中隐含的关键词，即机会，所形成的结论即是相关数据集的 Scenario map。此时，最重要的工作是对 Scenario map 进行分析和解释。以得出算法提取出的关键词对主体的影响作用，并采取相应地措施进行促进或抑制。因此，Scenario 分析是机会发现中非常重要的一个环节，分析结果的好坏直接影响到主体的决策。

目前常用的方法之一是利用概率求解，但是此方法只强调了如何区分 Scenario，而没有考虑到 chance 的表示以及管理问题。

文献[14]提出一种称为 smart data 的分析方法，在形成数据集的 KeyGraph 图后，根据图中获取的节点集合、边集合和桥集合进行分析，转换为逻辑表达式。并从中找出与 Scenario 有关的数据，在初始数据集中对该数据添加注释，

形成 smart data。该方法虽然可以有效地实现 Scenario 的比较和分析，但是该方法对于机会的表示形式描述模糊，且注释数据要依据操作者的经验设定，实现较为复杂。

3.4 机会发现模型

图 1 给出了机会发现过程的双螺旋模型[8]。在这个模型中，具有 2 个螺旋上升的子过程。一个是由人进行的机会发现螺旋上升过程。另一条螺旋线描述的是计算机的处理过程。双螺旋也就意味着这 2 个处理过程是并行的处理过程。即这些螺旋线的同时执行是基于输入的用于发现机会的目标数据而进行的。机会的确认最终是由人确定，这就规定了机会发现过程中人的核心地位。

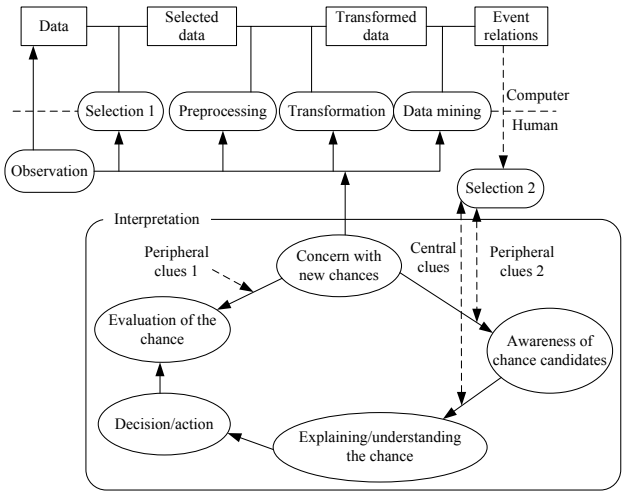

图 1 机会发现模型

4 未来研究方向

4.1 机会的形式化定义

虽然有很多学者从不同的角度对机会的定义进行研究并给出了相应的定义，但是仍没有形成统一的、普遍适用于大多数领域的机会定义。

机会首先是一个事件或状态，而事件/状态如何描述，如何给出能够体现机会的不确定性等特征的形式化定义，计算机内部表示形式如何定义，相应的转换等问题都是有待于进一步研究和解决的问题。

4.2 适用于大规模数据集的提取算法

信息可视化是机会发现过程中计算机系统的另一个重要任务。由于机会发现的数据集并不是具有很好的组织形式，且有大量的潜在结构。因此，机会发现的信息可视化不仅要可视化它自己，而且要给出数据特有的结构。目前主流的发现算法都是针对一篇文档而进行关键词提取的算法。因此，在面向大规模文档数据集时，算法的执行效率非常低。尤其是当数据来自网络时，在机会发现的初始数据集中，由于各局部数据源的数据模式是由不同的用户，在不同的时间和地点，基于不同的数据模型独立地设计的，它们之间可能存在着各种差异和冲突。

4.3 机会判定

机会发现是一个强调人机交互的过程。在此过程中，人的作用是至关重要的。发现过程中的一系列控制参数都要依据人的经验来确定，发现的结果的准确性主要取决于指导者的经验是否丰富，并且带有指导者强烈的主观愿望。因此，如何判断所提取的关键字(即机会)是否准确是文本机会发现研究的另一基本问题。

5 结束语

目前，随着信息技术的发展，文本成为信息的主要载体，利用文本机会发现方法，及时有效地发现隐藏于文本信息中的机会，对商业、经济、教育以及个人都具有极其重要的指导作用。在信息科学领域，文本机会发现是人工智能与数据库技术相结合的一个新兴的研究领域。本文从文本机会发现的基本理论研究出发，对文本机会发现的研究现状进行了深入的探讨，并给出了目前文本机会发现研究中仍迫切需要解决的问题。

参考文献

[1] Ohsawa Y. Introduction to Chance Discovery[J]. Journal of Contingencies and Crisis Management, 2002, 10(2): 61-62.

[2] 褚世卓, 陈小平. Chance Discovery 研究综述[J]. 计算机科学, 2004, 31(2): 1-5.

[3] Ohsawa Y. Dtection of Earthquake Risks with KeyGraph[M]// Ohsawa Y, McBurney P. Chance Discovery. [S. l.]: Springer Verlag, 2003: 339-349.

[4] Goda S, Ohsawa Y. Estimation of Chain Reaction Bankruptcy Structure by Chance Discovery Method with Time Order Method and Directed KeyGraph[J]. Journal of Systems Science and Systems Engineering, 2007, 16(4): 489-498.

[5] Lilia A, Tsang E P K. The Repository Method for Chance Discovery in Financial Forecasting[C]//Proc. of KES'06. Bournemouth, UK: [s. n.], 2006: 30-37.

[6] Ohsawa Y. Modeling the Process of Chance Discovery[M]// Ohsawa Y, McBurney P. Chance Discovery. [S. l.]: Springer Verlag, 2003: 2-15.

[7] Abe A. The Role of Abduction in Chance Discovery[J]. New Generation Computing, 2003, 21(1): 61-71.

[8] Ohsawa Y. Genetic Words Carrying Sequential Context[C]//Proc. of the 2nd International Workshop on Chance Discovery. Tokyo, Japan: [s. n.], 2002: 1-10.

[9] Ohsawa Y. KeyGraph: Automatic Indexing by Co-occurrence Graph Based on Building Construction Metaphor[M]//Ohsawa Y, McBurney P. Chance Discovery. [S. l.]: Springer Verlag, 2003: 262-275.

[10] Ohsawa Y. KeyGraph: Visualized Structure Among Event Clusters[M]//Ohsawa Y, McBurney P. Chance Discovery. [S. l.]: Springer Verlag, 2003: 262-275.

[11] Ohsawa Y. Chance Discovery with Data Crystallization: A Basic Research for Discovering Unobservable Events[J]. New Mathematics and Natural Science, 2005, 1(3): 373-392.

[12] Yutaka M, Ohsawa Y. KeyWorld: Extracting Keywords from a Document as a Small World[C]//Proc. of the 4th International Conference on Discovery Science. Washington D. C., USA: [s. n.], 2001.

[13] Mcburney P, Parsons S. Chance Discovery and Scenario Analysis[J]. New Generation Computing, 2003, 21(1): 13-22.

[14] Takanma Y, Iwase Y. Scenario to Data Mapping for Chance Discovery Process[J]. Soft Computing, 2007, 11(8): 773-781.

编辑 金胡考