摘　要

当今中国处于经济转型升级的关键时期，社会主要矛盾发生了历史性变化，涌现出诸多恶劣社会风险事件。Web2.0 时代提供了由用户主导而生成内容（UGC）的 Web 平台，这些风险事件在 BBS、微博上一经曝光，便迅速引起全社会的广泛关注，为政府管理带来了巨大挑战。搜索引擎承载了亿万网民的关注，是一个综合集成研讨厅实例，也是真实的舆情创造场。网民自发的、不受时空限制的对各种社会现象、社会问题进行搜索，是公众社会关注的真实体现。从这些海量的 UGC 数据中获取公众感知的社会风险，寻找结构化的方法描述社会风险事件，是复杂社会舆情问题结构化求解的过程，需要综合集成技术的支持。本文选取百度“热搜新闻词”做为研究社会风险感知的数据源，从多视角探索社会风险事件的结构化分析方法。通过一系列不同的计算抽取社会风险事件要素，进一步可视化社会风险事件演化进程及风险变迁过程；定量描述社会风险感知数据并探索其与社会、经济指数之间的关系。本文可看做从定性到定量综合集成方法论在社会风险分析中的成功运用。

本文主要研究内容包括：

1）从多学科角度介绍了社会风险研究的相关研究，进一步介绍了综合集成研究小组在基于互联网数据感知在线社会风险方面开展的一系列研究。通过探索社会风险事件结构化的分析方法将非结构的社会风险问题结构化。

2）定义 5W 抽取框架对风险事件进行结构化表示。5W 分别为地点（where）、时间（when）、人物（who）、原因（why）和内容（what）。将 5W 自动抽取转化为不同的机器学习任务，并根据具体任务提出并构建模型。引入条件随机场模型对热搜新闻词进行风险标注，有效提高风险标注精度。基于模型状态特征构建风险特征词库并分析各风险类别下的地域特征；基于热搜词新闻和风险特征词库抽取代表风险事件内容的风险关键词；基于地名抽取结果探索社会风险在时间和空间维度上的变化。

3）可视化表达社会风险事件随时间的演化过程。在 metro map 模型基础上进行扩展并提出在事件级别进行分析的 risk event map 模型，通过引入词嵌入模型学习风险事件的向量表示，进一步基于聚类算法获取事件簇；构造并优化目标

函数从而自动生成描述复杂社会风险事件演化过程的故事线图。以“红十字会和郭美美”事件为例进行分析，绘制“郭美美和红十字会”故事线的演化进程，并探测了风险随时间的变迁。

4）探索社会风险感知指标与社会、经济指数之间的关系。基于社会风险事件风险标注结果生成社会风险水平，基于格兰杰因果检验探索社会风险水平与百度指数之间的关系，并分析其对股票指数的影响，进一步考虑周末和节假日效应提出处理非交易日数据的建模方法。结果显示，百度搜索指数和社会风险水平之间存在格兰杰因果关系；金融经济、社会稳定和政府执政风险类对上证指数和深证指数的波动有显著影响，并验证了周末和节假日效应。以上结果表明基于公众在线搜索数据进行定量社会风险感知的可行性，以及社会风险水平可作为衡量社会运行状态的一种指标的有效性，为研究社会风险提供了一种新的分析角度。

关键词：社会风险，热搜新闻词，综合集成，5W，risk event map，格兰杰因果检验

Abstract

Modern China is at a critical period of social and economic transformation with the historical changes in the principal contradiction of Chinese society, emerging tremendous wicked societal risk events. In Web2.0 era, online platforms provide a variety of user generated contents (UGC). Those societal risk events expose to the public via BBS posts and microblogs, then quickly attract widespread attention and bring great challenges for government management. The search engine carries the attention of hundreds of millions of netizens. It is an example of hall for workshop on meta-synthetic engineering (HWMSE), as well a real creation ba of public opinion. Without being restricted by time and space, netizens spontaneously search for various societal phenomena and problems, truly reflecting public's concern. Exploring a structured approach to describe societal risk events is a solving process of complex societal problems when we perceive societal risk from these massive UGC data. Meta-synthesis technology is required. This paper chooses Baidu "hot news search words (HNSW)" as the data for study of socital risk perception. We explore the structured method of societal risk events from multiple perspectives. The societal risk event elements are extracted through a series of computing. Furthermore, the evolution process of societal risk event and transitions of risks along the time are visualized as well. Societal risk perception is quantitatively described and relationships with stock indexes are explored. This paper could be regarded as a successful application of the methodology from qualitative to quantitative meta-synthesis in societal risk analysis.

The main contents of this paper include:

1) The development of societal risk perception is reviewed. A series of studies of online societal risk perception which are conducted by research group of meta-synthesis and knowledge science are further introduced. Exploring the effective methods is a way to structure the unstructured societal risk problem.

2) To get a structured view of societal risk event, an event extraction framework

5W is proposed. 5W includes 5 elements, namely where, who, when, why, and what (5W). The tasks for extracting 5W of risk events are converted into different machine learning tasks. Effective approaches are put forward according to the tasks. Conditional random fields (CRFs) is applied for the risk category classification and achieves better performance. Its state features are picked out as risk factors to construct societal risk lexicon. The regional characteristics to each risk category are studied. Extract risk labeled keywords automatically which represent what the event occurred in a certain risk category. Finally，detect the changes of societal risks in time and space based on the results of geographical names extraction.

3) To study how societal risk events evolve along the time, this paper presents an improved algorithm “risk event map” which is processed at event-level based on the algorithm of “metro map”. Word embedding model that learns vector representation for societal risk events and clustering algorithm are adopted to process clusters. Construct and optimize the objective function to automatically generate storylines which build comprehensive views to describe the evolution of societal risk events. One case of “Guo Meimei and Chinese Red Cross” event is studied to explore the evolution process and analyze the transitions of risks along the time.

4) The relationships between societal risk perception and social and economic indexes are explored. The societal risk level is generated by tagging results of societal risk events. The relationship between societal risk level and Baidu index is explored based on Granger causality test, and the effects on stock index is analyzed as well. The weekend and holiday effects in China stock market are taken into consideration and methods for dealing with no-trading days data are proposed. The results show that there exist causal relations between Baidu Index and societal risk perception. The perception of finance & economy, social stability, and government management has distinguishing effects on the volatility of both Shanghai Composite Index and Shenzhen Composite Index. The weekend and holiday effects of societal risk perception on the stock market are verified. The research demonstrates that capturing societal risk based on on-line public concerns is feasible, as well as the validity of societal risk levels regarded as an

index to measure the state of social operation, providing a new perspective to conduct societal risk studies.

Key Words: Societal risk, HNSW, meta-synthesis, 5W, risk event map, Granger causality test