琉球大学学術リポジトリ

Unsupervised Classification for Main Features Extraction in Natural Disaster Text Sources

メタデータ	言語: 出版者: 琉球大学 公開日: 2015-04-14 キーワード (Ja): キーワード (En): 作成者: エンリケ, グティエレツ　カルロス メールアドレス: 所属:
URL	http://hdl.handle.net/20.500.12000/30654

Form 3

論　文　要　旨

Abstract

論 文 題 目

Title **Unsupervised Classification for Main Features Extraction in Natural Disaster Text Sources**

自然災害テキスト情報源の主要記事抽出に対する教師なし分類化に関する研究

After a natural disaster and during the post-disaster reconstruction time, on-line newspapers, social networks and blogs become very active describing many events interrelated. This work proposes several unsupervised models to extract main features and patterns from high dimensional text data generated during natural disasters. The main idea is to provide automatic and independent analysis tools of complex data which can be used rapidly when it is most needed. Firstly, we explore dimensionality reduction and patterns discovering by principal components analysis; an evolutionary description of news through the time is proposed by showing the activated principal components. Secondly, spatial and temporal properties of a news data set are extracted by self organizing maps (SOM). A model is proposed to obtain quantization points; these new vectors are clustered on map by K-means algorithm to detect potential patterns. Temporal dependency is detected by tracking SOM units activation over the time by a time-dependency matrix. Besides that, a linear prediction model is proposed to discover trend topics on news stream by uncovering the most influential variables. Each input is classified on the fly within 2 dummy categories and entered into a linear model with shrinkage operators where strongest variables prevail while those with negligible characteristics are removed. In addition, a random forest model composed for more than 200 decision trees is proposed to uncover predominant features from a large set of tweets, features are organized as a hierarchy of main variables where rules and an approximation of how information flows during an emergency are detected. Furthermore, particle filtering is applied to track a set of related words, a defined topic, within a news stream. Topic' relevance is estimated through the time by using a state-space model based on uni-gram model. Sampling importance re-sampling (SIR) algorithm is used to compute the posterior distribution value using available observations. Finally, a Bayesian model called Latent Dirichlet Allocation (LDA) is adapted to discover topics on Twitter stream text data, uncovering natural disaster related topics over the time; the inferences expose the concept of potential significant issues on real time.

Name GUTIERREZ, CARLOS ENRIQUE