Chapter 66

Privacy Preserving Text Analytics:
Research Challenges and Strategies in Name Analysis

Suresh Veluru
City University London, UK

Yogachandran Rahulamathavan
City University London, UK

B. B. Gupta
National Institute of Technology Kurukshetra, India

Muttukrishnan Rajarajan
City University London, UK

ABSTRACT

An e-mail address is a source of communication for major social networking sites. In general, e-mail addresses hold identity in the form a surname as a substring in it. Identities such as names are far from random and can exhibit community distributions over populations. However, these identities reflect cultural, ethnic, and genetic structures generated among populations. Hence, identity establishment in e-mail address mining can be seen as a categorization of e-mail address-based community structure in names data set. It involves community modeling in names, categorization of an e-mail addresses, and identity privacy preservation. This chapter presents a survey of text mining and privacy preserving techniques followed by research challenges and strategies in name analysis. The research challenges are: (1) e-mail address categorization based on community structure of identities, (2) correlation of surnames and forenames within and across communities, and (3) privacy preserving of identities in communities.

DOI: 10.4018/978-1-4666-8111-8.ch066

Copyright © 2015, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

INTRODUCTION

Personal names practice exists in all human groups which have different distributions for different communities. Identities like forenames, surnames, and names represent ethnic, geographic, cultural and genetic structures that have developed in human populations. In recent years, it has been noticed that people are migrating from one location to another location due to job prospects, economic prosperity, political unrest or globalization. Indeed, the names of migrants retain semantic similarity to the names of the people at their original locations. Identities exhibit several distributions in which each distribution is a community. These distributions may or may not overlap. Initial intuitions of community modeling are i) identities at each location have proportion of several communities, and ii) identities collected at several locations contain mixture of community distributions. These communities can be structured hierarchically. For example, intentional Christian communities are as diverse as the communities' movement and the many kinds of Christianity. Many are inspired by biblical passages to be Christian communes. Others have separate finances and might be similar to co-ops, land trusts, or co-housing. Identity establishing in e-mail address mining can be seen as an information retrieval in names and e-mail address data sets.

Information retrieval in text has great potential which would turn text documents into useful information. Statistical information of terms is represented in a vector space model to extract information in a set of documents (Aas & Eikvil, 1999; Manning & Schutze, 2003). Perhaps, phrase based approaches do not perform well since phrases do not repeat as the terms repeat in a set of documents. Hence, statistical information of terms is useful to extract knowledge effectively (Sebastiani, 2002). Similarly, surnames provide good statistical information at each location which can be used in vector space model to extract information in names data set (Veluru et al., 2013; Veluru et al., 2012).

Names analysis have been developed in geography such as identifying spatial concentration of surnames (Cheshire & Longley, 2011), migrant surname analysis (Longley, Cheshire, & Mateos, 2011), uncertainty in the analysis of ethnicity classification (Mateos, Singleton, Longley, 2009), and ethnicity and population structure analysis (Mateos, Longley, & O'Sullivan, 2011). Naming network concept has been developed by Mateos et al. (Mateos, Longley, & O'Sullivan, 2011) which uses a large graph to represent relations among forenames and surnames to detect community structure. Isonymy (Rodriguez-Larralde et al., 1994) and Lasker distance (Lasker, 2002) are used which define the degree of similarity between surname mixes by comparing relative frequencies of surnames at different locations. These measures are complementary measures such that the inverse natural logarithm of the isonomy creates a more intuitive measure called Lasker distance. These are applicable to study inbreeding between marital partners or social groups, but do not explicitly address the semantic similarity between surnames.

An e-mail address often contains an identity as substring in it. Identity establishment in e-mail address mining establishes community identification from an e-mail address if an e-mail address contains identity as a substring. Hence, e-mail address categorization can be done based on hidden community structure in identities. It has many applications such as to predict individual community from his (or her) e-mail address or social network profile. It can predict if two or more individuals belong to same community (or not) from their e-mail addresses. It can be used to classify e-mail addresses based on ethnic, geographical and genetic structure among people. We have applied vector space model, latent semantic analysis and performed e-mail address categorization based on semantics of surnames (Veluru et al.2013). Further, privacy preserving

19 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/privacy-preserving-text-analytics/125352

Related Content

Profiles and Motivations of Standardization Players
Cesare A. F. Riillo (2013). *International Journal of IT Standards and Standardization Research (pp. 17-33).*
www.irma-international.org/article/profiles-and-motivations-of-standardization-players/83545

Community-Driven Specifications: XCRI, SWORD, and LEAP2A
Scott Wilson (2010). *International Journal of IT Standards and Standardization Research (pp. 74-86).*
www.irma-international.org/article/community-driven-specifications/46114

Trends in Information Security
Partha Chakrabortyand Krishnamurthy Raghuraman (2015). *Standards and Standardization: Concepts, Methodologies, Tools, and Applications (pp. 1582-1604).*
www.irma-international.org/chapter/trends-in-information-security/125359

The Implications of Alireza Noruzi’s Laws of the Web for Library Web-Based Services
Josephine Eruterio Onohwakporand Benson Oghenevwogaga Adogbeji (2011). *Handbook of Research on Information Communication Technology Policy: Trends, Issues and Advancements (pp. 724-733).*
www.irma-international.org/chapter/implications-alireza-noruzi-laws-web/45420

Should Buyers Try to Shape IT Markets Through Non-Market (Collective) Action? Antecedents of a Transaction Cost Theory of Network Effects
Kai Reimersand Mingzhi Li (2005). *International Journal of IT Standards and Standardization Research (pp. 44-67).*
www.irma-international.org/article/should-buyers-try-shape-markets/2563