Text Data Mining in a Geospatial Metadata Catalog

Fátima Ferrão dos Santos*
Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, Brazil – fátima.santos@ibge.gov.br

Abstract

The characteristics that make geospatial data special as a computing problem have been acknowledged in many ways and one of the mainly important initiatives is the development of Spatial Data Infrastructure (SDI). This term refers to the integrated set of technologies, standards, policies, institutional arrangements and human resources to facilitate the availability, access and use of geospatial data and information to support policy, business, research, govern and society at large. In order to share the geospatial information, it is necessary to deploy metadata. Mining metadata can enable the extraction of patterns and useful and new knowledge in the context of this application. The aim of this paper is to present a methodology to implement text data mining in a geospatial metadata catalog. The focus of this paper is not on developing the spatial data infrastructure but on developing data mining tasks within the emerging infrastructure. To assess this information, a network approach based in probability estimative was used. These techniques are aimed at discovering natural divisions of networks into groups, based on metrics of strength of connection between vertices. Comparisons among the generated networks are made in the light of different categories of metadata, the closeness and betweenness centralities measurements. Our results show the main pattern among distinct categories and how keywords are correlated to these main categories. We conclude that the approach of complex networks is a promising tool for metadata analysis.

Keywords: metadata; text data mining; spatial data infrasctructure; networks.

1. Introduction

The development of technologies as the Internet, general information systems, cost reduction of data storage, remote sensing and survey technologies, over the last years, has dramatically enhanced our capabilities to collect terabytes of geographic data on a daily basis. However, the ability to analyze these data, turning them into useful knowledge is much less than the production capacity and storage. This confronts us with an urgent need for new methods and tools to aid the man in the task of analyzing, interpreting and relating those data, transforming them in useful knowledge and making possible the development of action strategies in the context of their application. Data Mining research has been considered to describe the overall process that empowers the abilities to extract new, insightful information embedded within large heterogeneous databases and to formulate knowledge.

In Brazil, the SDI gets the name of the Infraestrutura Nacional de Dados Espaciais (INDE), established by Decree 6666 of 27/11/2008 in order to catalog, integrating and harmonizing geospatial data produced or kept by institutions of various types, especially the government, providing their dissemination and use. With such information available for services provided via the Internet, by public international protocols, widely accepted, it is possible to locate and access them more simple, fast, comprehensive and integrated way, and no specialized knowledge is required. One of the main objectives is to avoid duplication of efforts and waste of resources to obtain geospatial data by public administration, through the dissemination of metadata relating to such data available on the entities and public bodies at all levels of government.

Metadata is a critical component of any spatial data infrastructure initiative. It not only provides users of spatial data with information about the purpose, quality, actuality and accuracy and many more of spatial datasets, but also metadata performs crucial functions that make spatial data interoperable, that is, capable of being shared between systems. Mining metadata in the context of SDI is a powerful tool to identify patterns in the data, as well as anomalies and errors in filling in metadata.

Several studies have shown how the structural characteristics of networks are related to their stability and dynamic (Albert et al., 2000; Albert et al., 1999), showing how some features followed well-defined patterns explained by network theory. So, the aim of this paper is (i) to apply data mining tasks to identify interdependence between sets of categories and keywords in metadata (ii) to use a network approach to investigate possible wiring patterns (iii) apply the closeness and betweenness centralities measures in order to determine the positional importance of each keyword or node; (iv) identify which set of n nodes belongs to the core network for each distinct metadata category. The outline of this paper is as follows. This introduction is the section 1. In section 2, we describe the materials and methods: data, indexes applied and the concepts behind the implementation of our methods for finding community structure. In section 3 we present results and, in section 4, the discussions. Section 5 presents the conclusions.

2. Material and Methods

2.1 Available Data and Software

The samples were presented in separate measurements between groups of categories and keywords considered in our geospatial metadata catalog. The total number of valid terms considered: (i) forty-five category terms and (ii) one hundred and three of keyword terms. Cases where multiple keyword terms are stored on a single keyword attribute of their metadata were not considered valid.

To perform the analysis, which has not been completed, the following tools were used: (i) xml - services to read the catalog content and obtain the necessary metadata in XML format; (ii) text data mining tasks to extract the frequencies of category and keywords from the xml files.

2.2 Network generation

In this paper we present an approach for the discovery of community structure in networks with only a single type of vertex and a single type of undirected and unweighted edge, although generalizations to more complicated network types are possible. Each vertex is a word (category or keyword). Many studies of networks have been the subject of recent research, mainly the identification of community structure, through the division of network nodes into groups within which the networks connections are dense, but between groups are sparse. In this paper we use the approach of interaction network analysis to discover natural divisions of networks into groups. A well-known interdependence technique is a clustering process. Some networks do not have natural metrics, but suitable ones can be devised using correlation coefficients, path lengths, or matrix methods. The procedure can be halted any point, and the resulting components in the network are taken to be the communities. In a divisive method, we start with the network of interest and attempt to find the least singular connected pairs of vertices and then remove the edges between them. Applying this method repeatedly, we divide the network into smaller and smaller components. Likewise, we can stop the process at any stage and take the components at that stage to be the network communities. There is a wide variety of methods to identify when to stop the process, independently if it is an agglomerative or divisive one. Our algorithm is a divisive method with some peculiarities. Rather than looking for similarities between vertices, we will look for the edges in the network that are most frequently (and with more intensity) linking two or more vertices. So, our divisive algorithm focus is not on removing the edges between vertex pairs with low similarity, but on finding edges with the highest values of occurrences, i.e. we focus on finding community structure based on the values of the edges and not on the attributes of the vertices, as is more usual. For each simultaneous occurrence between a vertex pair, there is an edge with an interaction value. To quantify these interactions, it was considered the probability of a variable Bi given the presence of a variable Bj, represented by P (Bi|Bj), which measure the strength of the association between Bi and Bj. As P (Bi|Bj) does not account statistical confidence, we consider the equation 1, proposed by Stephens et. al. (2009), as a measure of confidence.

$$e\ (Bij \mid Bj) = NBj\ (P(\ Bi \mid Bj) - P(\ Bi)) / (NBj\ (P(\ Bi)\ (1 - P(Bi)))\ ½ \qquad (1)$$

Essentially, it is a one-sided binomial test where the null hypothesis is that the distribution of Bi is random over the data collected.

2.3 Applied indices

The topology of a network depends on the interaction between nodes and the importance of each node is crucial to determine their dynamics and stability (Strogatz, 2001). For Freeman (1979), the importance of each node is usually quantified using centralities indices. However, the choice for using a specific index depends on the type of information we want to get. Different indices reflect different aspects about the importance of a node within a network (Jórdan, 2009). We applied two of the most common measures of centrality: closeness and betweenness. Both measures express the concept of distance (i.e., the shortest paths) between nodes, consequently, similarity among networks. The closeness centrality of a vertex is based on the total distance between one vertex and all other vertices, where larger distances yield lower closeness centrality values. As closer the vertex to all others as higher it centrality is. In the network theory, it is defined as the mean geodesic distance between a vertex v and all other vertices reachable from it, such as equation 2:

$$\frac{\sum\limits_{t\in V\setminus v} d_G(v,t)}{n-1} \qquad (2)$$

Where n >= 2 is the size of the network's connectivity component V reachable from v. Closeness can be regarded as a measure of how long it will take information to spread from a given vertex to other reachable vertices in the network (Newman, 2003). The centrality of a vertex depends also on the extent to which it is needed as a link in the chain of connections within the network. According Nooy et. al. (2005), geodesics is the shortest distance between two nodes in a network. If we consider the geodesics to be the most likely channels to facilitate aggregation between variables, the one that is situated in the geodesics between many pair of vertices is very important in the network. This approach is based on the concept of betweenness. The betweenness centrality of a vertex is the proportion of all geodesics between pairs of other vertices that include this vertex and measures how crucial a node is for the exchange of information within a network. Betweenness is a centrality measure of a vertex within a graph so that, vertices that occur on many shortest paths between others have higher betweenness than those that do not. For a graph G: = (V,E) with n vertices, the betweenness CB(v) for vertex v is computed as follows:

1. For each pair of vertices (s,t), compute all shortest paths between them.
2. For each pair of vertices (s,t), determine the fraction of shortest paths that pass through the vertex in question (here, vertex v).
3. Sum this fraction over all pairs of vertices (s,t), such that equation 3 (Shivaram, 2005:

$$C_B(v) = \sum_{s\neq v\neq t\in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \qquad (3)$$

where σst is the number of shortest paths from s to t, and $\sigma st(v)$ is the number of shortest paths from s to t that pass through a vertex v. This may be normalised by dividing through the number of pairs of vertices not including v. Closeness and betweenness indexes are based on the position of the node in the network.

3. Results

We identified patterns through the analysis of the networks. We want to measure not only the interaction, but the intensity of it. We calculated the value of each edge (equation 2) for each pair category-keyword. The calculation of the summation value of an edge (all samples) is simple to implement. After that, we calculated the frequency distribution of edges over the values. We applied our proposed procedure to identify the core network of each category. Thus, the general form of our network core structure finding algorithm is as follows:

1. Calculate the value of each edge (equation 2).
2. Calculate the sum of the values of the edges.
3. Apply clustering algorithms considering the values of the edges.
4. Consider cluster which has the highest values of edges, the core network.

Finally, after step 4, for each distinct pair of clusters, identify the edge (if there is one) with the highest value that lie between them and consider it as their main connection. This point differentiates our approach from clustering algorithms, but do not invalidate its use and implementation. In cluster analysis, the attempt is to maximize the homogeneity of objects within the clusters while also maximizing the heterogeneity between the clusters. In our case, the identification of the edge with the highest value between sub-networks, after the division process, is based on the concept that the edge with the highest value between sub-networks is more important for the spread of information between them. With this process, it is possible to identify the synchronized components of the resulting sub-networks, i.e. the components that occur simultaneously. We identified the way metadata is being filled by the different actors in our SDI. Figure 1 shows how categories and keywords of our metadata catalog link to each other. The analysis of this network proved that the appropriate terms has been used. It was also possible to identify distinct types of errors and mainly it was possible to have insights to propose a domain-level controlled vocabularies.

4. Discussion

Following the key role of metadata in an SDI, the existence of metadata has been acknowledged as one the fundamental indicators for assessing any spatially enabling platform. In this regard, to achieve a spatially enabled platform delivering complete, and reliable metadata to end users through the spatial data catalog system has been a concern of data providers.

We identify the need for data infrastructures, which we contend, will need to be powered by semantic technologies. These are needed to provide computational approaches that will allow users to search for and discover data resources, rapidly integrate large-scale data collections from heterogeneously collected resources or multiple data sets, and compare these results across datasets to allow generation and validation of hypotheses. To move to more integrative capabilities, we need to adapt existing infrastructures, implement ways to link them, and develop means to accommodate the aforementioned complexities: scale, scarcity, model, and uncertainty. In attacking these latter issues, there needs to be more widely deployed infrastructure for sharing, and preserving data. This comes about for a number of reasons, particularly reducing duplication and encouraging actors to know more about each other's results. There are also other reasons that are more specific to data. One of the reasons is that people are discovering that data collected for one reason can often, later, be used for another purpose, especially when even simple cross-correlations can be performed. One key aspect of this work that is identifying ways that metadata and geospatial information can be found and used. For this purpose, we need to develop domain-level controlled vocabularies (or ontologies) that can be used to navigate through INDE.

5. Conclusions

The main contribution of this paper is to show how the representation of text interaction could be constructed through a network approach to discriminate those terms of greater influence in our SDI. It was possible to identify the core network of geospatial information categories and keywords, and their similarities. In the studied case the properties of the corresponding network show to what extent a given category of information is exploiting its potential through keywords. The application of complex networks approach proved to be an useful tool for semantic diagnostic. Although significant results were obtained, it is necessary to study deeply other scenarios. This work is already being developed.

References

Albert R, Jeong H. Barabási AL, 2000, Error and attack tolerance of complex network. Nature 406: 378-382.

Albert R, Jeong H. Barabási AL, 1999, Diameter of world-wide web. Nature 401: 130-131.

Freeman L, 1979. Centrality in social networks: conceptual clarification. Social Networks I: 215-239.

Jordan F, 2009. Keystone species and food webs. Philosophical Transactions of the Royal Society of London Series B – Biology Sciences 364: 1733-1741.

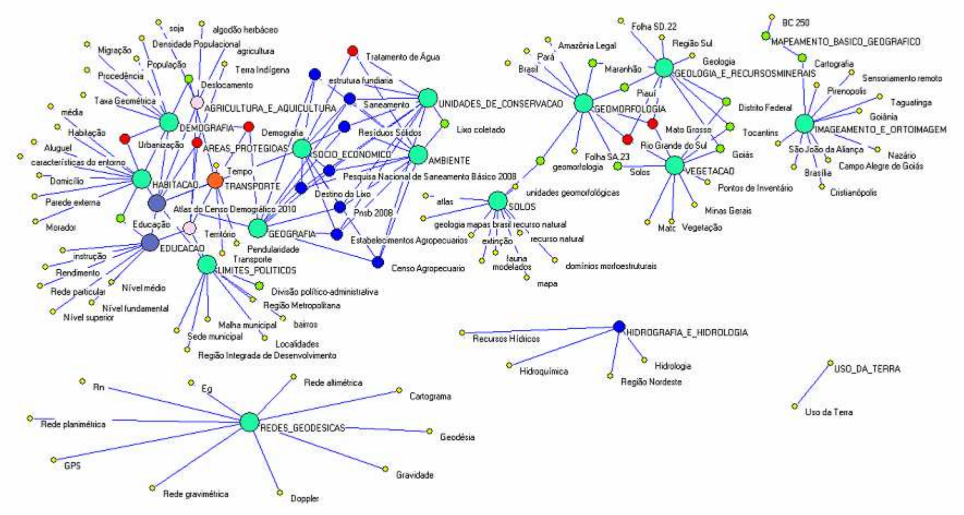
Nooy W., Mrvar A., Batagelj V. 2005. Exploratory Social Network Analysis With Pajek. Structural Analysis in the Social Sciences 27, Cambridge University Press.

Stephens CR, Heau JG, Gonzalez C, Ibarra-Cerdenã CN, Sanchez-Cordero V, et al. 2009, Using Biotic Interaction Networks for Prediction in Biodiversity. and Emerging Diseases. PLoS ONE 4(5): e5725. doi:10.1371/journal.pone.0005725

Strogatz SH, 2001. Exploring complex networks. Nature 420: 268-276.

[USCOP] US Commission on Ocean Policy. 2004. An Ocean Blueprint for the 21st Century.Washington (DC): USCOP.

Figure 1. Interaction Network of Categories and Keywords



soja
algodão herbáceo
Migração
Densidade Populacional
agricultura
População
Terra Indígena
Procedência
Deslocamento
Taxa Geométrica
AGRICULTURA_E_AQUICULTURA
média
DEMOGRAFIA
Demografia
Habitação
Urbanização
AREAS_PROTEGIDAS
Aluguel
características do entorno
Tempo
Domicílio
HABITACAO
TRANSPORTE
Parede externa
Atlas do Censo Demográfico 2010
Morador
Educação
Território
EDUCACAO
Pendularidade
Transporte
instrução
Rendimento
LIMITES_POLITICOS
Rede particular
Nível médio
Nível fundamental
Nível superior
Divisão político-administrativa
Região Metropolitana
Malha municipal
bairros
Sede municipal
Localidades
Região Integrada de Desenvolvimento
Tratamento de Água
estrutura fundiaria
Saneamento
UNIDADES_DE_CONSERVACAO
Resíduos Sólidos
Lixo coletado
SOCIO_ECONOMICO
AMBIENTE
Pesquisa Nacional de Saneamento Básico 2008
Destino do Lixo
Pnsb 2008
GEOGRAFIA
Estabelecimentos Agropecuarios
Censo Agropecuario
Folha SD.22
Amazônia Legal
Região Sul
Pará
Brasil
Maranhão
Geologia
GEOLOGIA_E_RECURSOSMINERAIS
Piauí
GEOMORFOLOGIA
Distrito Federal
Mato Grosso
Rio Grande do Sul
Tocantins
Folha SA.23
geomorfologia
Solos
VEGETACAO
Goiás
unidades geomorfológicas
Pontos de Inventário
atlas
SOLOS
Minas Gerais
geologia mapas brasil recurso natural
Mata
Vegetação
extinção
recurso natural
fauna
modelados
domínios morfoestruturais
mapa
BC 250
MAPEAMENTO_BASICO_GEOGRAFICO
Cartografia
Sensoriamento remoto
Pirenopolis
Taguatinga
Goiânia
IMAGEAMENTO_E_ORTOIMAGEM
São João da Aliança
Nazário
Brasília
Campo Alegre de Goiás
Cristianópolis
Recursos Hídicos
HIDROGRAFIA_E_HIDROLOGIA
Hidrologia
Hidroquímica
Região Nordeste
USO_DA_TERRA
Uso da Terra
Rn
Eg
Rede altimétrica
Cartograma
Rede planimétrica
REDES_GEODESICAS
Geodésia
GPS
Gravidade
Rede gravimétrica
Doppler