Ontology based Model and Procedure Creation for Topic Analysis in Chinese Language [*]

Dong Han and Kilian Stoffel

Information Management Institute, University of Neuchâtel
Pierre-à-Mazel 7, CH-2000 Neuchâtel, Switzerland
{dong.han,kilian.stoffel}@unine.ch

Abstract. This paper focuses on setting up a methodology to create models and procedures for the functions and processes involved in topic analysis, especially for the business documentation in the Chinese language. Ontologies of different types are established and maintained containing the annotated and evolved knowledge. Extraction, transforming and loading methods are adapted from approaches which are used to set up a data warehouse for standardized data models, exploited as the basis of a large variety of analysis. Topic discovery is conducted based on Latent Dirichlet Allocation for different usage. An interactive tool is implemented to support the proposed design and realistic demands.

Key words: Ontology, ETL, data warehousing, Chinese language, LDA

1 Introduction

For enterprises, a large number of documents are involved in the daily business and the size of the documents is also increasing with the pervasive usage of electronic forms. Therefore the motivation is high to facilitate their operation, not only to get an intuition of the content of the documents, but also to get a more profound comprehension of the semantics of the documents and the relationship between groups of keywords. If this objective can be achieved, then it will be much more effective for enterprises to integrate this type of operations into their processes of decision making, auditing, and market promotion. This idea, furthermore, can be adapted to a wide range of specific domains such as financial analysis, quality control, logistics management and many more. In many cases, latent topics are only implicitly presented in the documents. The classical way to handle this problem is to conduct statistical studies. This approach is particularly well suited for the cases in which a relatively small number of words are used very frequently. If, however, a larger number of words are distributed in the documents with latent links, then some novel approaches are required.

[*] This work is supported by Swiss National Science Foundation(SNSF) project "*Formal Modelling of Qualitative Case Studies: An Application in Environmental Management*" (Project No. CR21I2_132089/1).

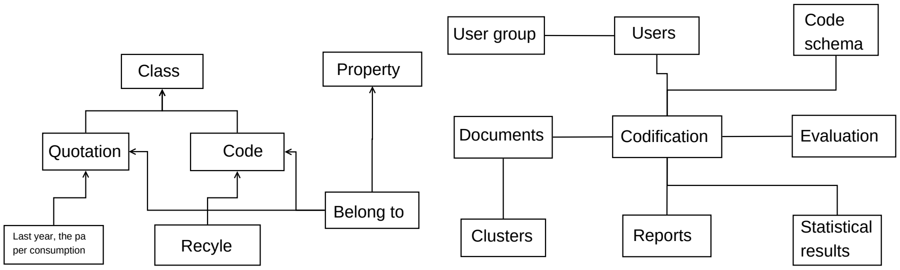
Meanwhile, as the economic development in China progresses, business based on Chinese language is getting influential in the global market. Documents written in Chinese are more difficult for traditional approaches which are based on the assumption of alphabets. We proposed a method to analyze the Chinese language in [1] with data mining approaches. As research advances, we have set up a more profound mechanism, leveraging the specific characteristics of Chinese language applied in different domains including financial management, enterprise management, etc.

The contribution and innovation of this paper is to propose a methodological framework by taking advantages of data driven approaches. It fully considers the re-usability of the existing systems from the point of view of the data, and avoids redundancy and repetition of functions. Domain strategies can also be estimated by applying this framework to verify their effectiveness. A prototypical tool has been implemented as a working platform to validate the proposed design. This prototype has already been deployed in real user cases to verify the effectiveness of the proposed methods.

2 Theory and Ontology Establishment

Considering the qualitative features of the data involved in this paper, we select *Grounded theory* [2] as the guideline to develop our new approach. This approach is appropriate for the research of Chinese documentation since there is usually plenty of data at the beginning of an analysis without conclusive statements. Inheriting from the principles of Grounded theory, three types of ontologies are set up: *Project ontology* is a high level ontological framework to define the objects and their relationships to ensure the compatibility. *Reference ontology* contains the incorporated knowledge from domain experts. *Code ontologies*, as depicted in figure 1(a), record the raw data from the primary texts and the user's annotation. OWL is utilized as the ontology language and depending on the concrete scenarios, we can use *lite*, *DL* and *full* standards.

Whilst generic ontologies are mainly designed for English and other alphabetical languages, ontologies based on the Chinese language are established considering its typical features. *Speech ontologies* represents the speech of the words of the original texts as well as the annotations. We use the speech ontologies here not aiming to interpret the sentences from one language to another. Instead, we would just like to comprehend the structures of the sentences for their subjects, verbs, objects, etc. This serves as the fundamentals of the further analysis. *Auxiliary ontologies* is a set of the bag of keywords which are highly influential to the grammar of the languages. They contain the major auxiliary and assistant words in the Chinese language, such as the words to express tense, tone, and meanings. Also the words for negation and interrogation are listed in these ontologies. *Localization ontologies* are set up to represent the knowledge about localization on top of the language per se. For example, the regional divisions in China include North, Central China, South with different dialects, terms, and expressions. Ontologies in this category are exploited with both language and domain knowledge to produce derived information. These ontologies play an important role in analyzing the documentation in Chinese language [1]. They also serve as the input of the ETL processes presented in the following section.



(a) Code ontology (b) Business data model

Fig. 1. Code ontology and business data model

3 ETL and Data Warehousing

In order to extract information from the ontological knowledge, we need to formulate a set of dimensions as the properties of this knowledge. For each ontology file, two tuples, external and internal respectively, are set up. All the extractions are supposed to be based on these tuples. Correspondingly, two kinds of extractions are designed - global extraction and local extraction. *Global extraction* aims at handling external information of the submissions made by experts, and is not going inside the files. The purpose of this step is to reorganize all the data in a structured way, labeling each file with its properties in accordance with the tuple previously defined. Once we finish this step, all the knowledge is stored in a uniform format, each file with its name containing the attributes in tuple 1. To facilitate this process of global extraction and to conform with industry standards, we use Powershell [3] scripts to carry out the extractions.

Next, *local extraction* is conducted. Local extraction aims to retrieve internal information from the ontologies following the definition of tuple 2. The results of this step are divided into two categories with respect to their data structures: (1) *Relational tables.* In a relational table, each attribute represents one dimension of the attributes, for example, company, year, code, code frequency. As a format well fitting the schema of relational databases, relational table is straightforward to be imported into a database management system for advanced queries. (2) *Pivot tables.* A second type of tables used in the system are pivot tables. A pivot table, used for data manipulation such as aggregation and sorting of the data, is mainly for the purpose of data output and graphics in tabular forms [4].

$$Q_{external} = \{group, user, project, document, year, file\} \tag{1}$$

$$Q_{internal} = \{file, quotation, coordinates, codes, ratings\} \tag{2}$$

With the two types of extractions, it is sufficient to extract and maintain most of the useful information from the ontologies. Moreover, subjects are derived from the business scenario, with data loaded via the ETL processes. These subjects are presented to depict a general view of the data involved in a project. Above that, a business data model is established to characterize the entities in each subject area in a finer granularity in order to reveal more details as described in figure 1(b). Furthermore data marts can be established and data mining methods can be carried out on top of these models.

4 Topic Analysis

The steps presented above facilitate the processes of topic analysis composed of two parts: data annotation and systematic analysis. Data annotation is carried out at the beginning by the users in interaction with the primary documents and recorded in the form of ontologies. This will provide refined data on top of the original texts. Once the annotation has been started, we need to extract the terms which are significant to the texts. For the terms, certain features can be extracted in respect to different dimensions of interest. With enough flexibility of feature construction, the main functionality to be provided during this step is the aggregation. For each document d, as an example, $(\mathbf{Q}_d, \mathbf{C}_d)$ is created as a matrix recording the quotations and their codes in this document. They are used as input of the term-topic model. There are several topic models which have been applied to discover the topics from documents. We propose to use LDA as it has advantages over other methods as shown in [5]. The formulas given by [6] illustrate the basic idea of LDA (see figure 2). One of the most important tasks for LDA is to estimate the parameters involved in the model, and the effectiveness of the parameter estimation highly influences the output of the topic discovery and analysis. A list of algorithms, such as the variational EM method [6] and the Gibbs sampling algorithm [7], are considered as the candidates to be leveraged to estimate the parameters involved in the LDA processes. The topics will then be generated based on the parameter estimation of these algorithms.

5 Implementation

A prototype, namely *Qualogier*, has been implemented as a working platform and test bed of the proposed approach (see figure 3). Based on ICEPdf [8], it takes PDF files as the primary documents, providing operations on these files such as turning pages, zooming in/out, printing, and extracting the texts. Users are able to select sentences and phrases from the texts of interest in the form of quotations and then assign codes to them. All the user behaviors are recorded in ontologies and then, together with the original texts, modeled as the background for establishing the data warehousing models. Furthermore different utilities are

```
1. Choose $N \sim Poisson(\lambda)$.

2. Choose $\theta \sim Dir(\alpha)$.

3. For each of the $N$ words $w_n$:

    (a) Choose a topic $z_n \sim Multinomial(\theta)$.

    (b) Choose a word $w_n$ from $p(w_n|z_n, \beta)$, a multinomial

probability conditioned on the topic $z_n$.
```

Fig. 2. LDA procedures (Source: David M. Blei, Andrew Y. Ng, and Michael I. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research 3 (2003), p. 996 ([6]))

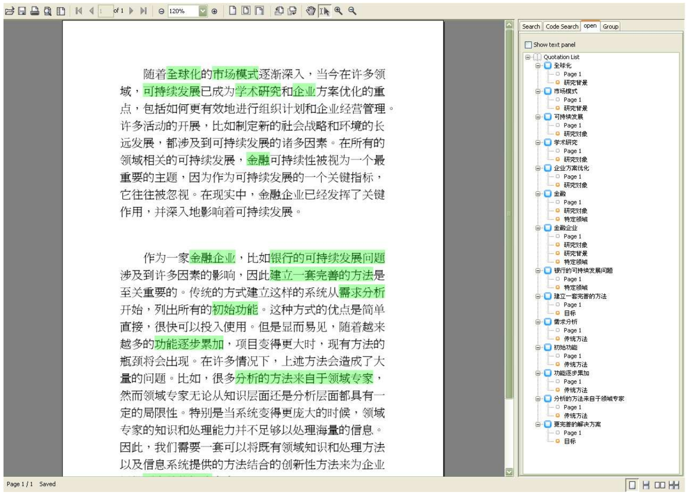
provided for exporting and outputting information using ETL methods. The topics generated by LDA are presented along with the original documents to show the key concepts based on their semantics and the users' annotations. The system also supports ontology inference based on forward chaining and backward chaining methods to produce new facts, for example, recommendations to the users. This system has been deployed in our research project for 40 domain experts to evaluate the sustainability performance of different firms in the form of a case study. These experts have various operating systems, domain knowledge, and research subjects. They are working with the proposed system in an interactive way as shown in figure 4. They select and highlight reports from the firms and submit them to the system. Based on the analytical system feedback, the experts will proceed with further studies. This system has several advantages compared to other similar systems like ATLAS.ti [9].

6 Conclusion and Future Work

In this paper, a methodology is presented based on ontologies, ETL, data warehousing, and LDA to retrieve information from the original data and model the entire working processes for documentations in the Chinese language. In the future, we will conduct more evaluation approaches and user experiments to verify the effectiveness of the methodology presented in this paper.

References

1. Han, D., Stoffel, K.: Ontology based Qualitative Methodology for Chinese Language Analysis. In: Proceeding of the 26th International Conference on Advanced Information Networking and Applications Workshops (WAINA). (2012)
2. Glaser, B., Strauss, A.: The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine (1967)



随着全球化的市场模式逐渐深入，当今在许多领域，可持续发展已成为学术研究和企业方案优化的重点，包括如何更有效地进行组织计划和企业经营管理。许多活动的开展，比如制定新的社会战略和环境的长远发展，都涉及到可持续发展的诸多因素。在所有的领域相关的可持续发展，金融可持续性被视为一个最重要的主题，因为作为可持续发展的一个关键指标，它往往被忽视。在现实中，金融企业已经发挥了关键作用，并深入地影响着可持续发展。

作为一家金融企业，比如银行的可持续发展问题涉及到许多因素的影响，因此建立一套完善的方法是至关重要的。传统的方式建立这样的系统从需求分析开始，列出所有的初始功能。这种方式的优点是简单直接，很快可以投入使用。但是显而易见，随着越来越多的功能逐步累加，项目变得更大时，现有方法的瓶颈将会出现。在许多情况下，上述方法会造成了大量的问题。比如，很多分析的方法来自于领域专家，然而领域专家无论从知识层面还是分析层面都具有一定的局限性。特别是当系统变得更庞大的时候，领域专家的知识和处理能力并不足够以处理海量的信息。因此，我们需要一套可以将既有领域知识和处理方法以及信息系统提供的方法结合的创新性方法来为企业

□ Show text panel

- Quotation List
  - 全球化
    - Page 1
    - 研究背景
  - 市场模式
    - Page 1
    - 研究背景
  - 可持续发展
    - Page 1
    - 研究对象
  - 学术研究
    - Page 1
    - 研究对象
  - 企业方案优化
    - Page 1
    - 研究对象
  - 金融
    - Page 1
    - 研究对象
    - 特定领域
  - 金融企业
    - Page 1
    - 研究对象
    - 研究背景
    - 特定领域
  - 银行的可持续发展问题
    - Page 1
    - 特定领域
  - 建立一套完善的方法
    - Page 1
    - 目标
  - 需求分析
    - Page 1
    - 传统方法
  - 初始功能
    - Page 1
    - 传统方法
  - 功能逐步累加
    - Page 1
    - 传统方法
  - 分析的方法来自于领域专家
    - Page 1
    - 传统方法
  - 更完善的解决方案
    - Page 1
    - 目标

Fig. 3. Screenshot of the system

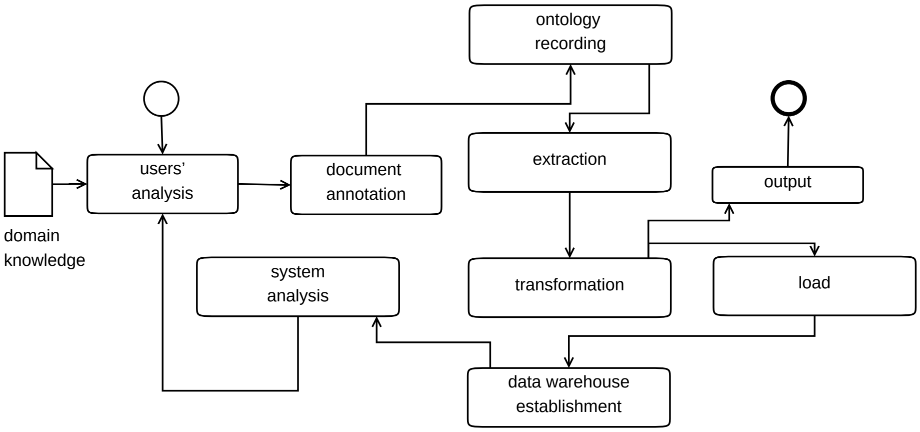


Fig. 4. Workflow of the entire processes

3. Microsoft: Powershell. http://technet.microsoft.com/en-us/library/bb978526.aspx
4. Wikipedia: Pivot table. http://en.wikipedia.org/wiki/Pivot_table
5. Gimpel, K.: Modeling Topics. Information Retrieval **5** (2006) 1–23
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3** (2003) 993–1022
7. Griffiths, T.L., Steyvers, M.: Finding Scientific Topics. In: Proceedings of the National Academy of Science. (2004) 5228 – 5235
8. Icepdf: Icepdf. http://www.icepdf.org/
9. Han, D., Stoffel, K.: An Interactive Working Tool for Qualitative Text Analysis. In: Proceeding of the 12th Francophone International Conference on Knowledge Discovery and Management. (2012)