Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 91 (2016) 452 – 459
Procedia Computer Science

Information Technology and Quantitative Management (ITQM 2016)

Combining Channel Theory, HowNet and Extension Model to Analyze Big Data

Weihua Li[a]*, Chunyan Yang[b]

[a]*School of Computer, Guangdong University of Technology, Guangzhou, 510006 China*
[b]*Research Institute of Extenics and Innovation Methods, Guangdong University of Technology, Guangzhou, 510006 China*

Abstract

Because the diversity of unstructured data has brought new challenges to big data analysis, this paper proposes to combine Channel theory, HowNet and extension model to improve big data analysis ability. The paper proposes a new method to process big data, which is based on the Channel theory idea and HowNet structure, in order to overcome the semantic conflicts of big data. In view of the problems that people are difficult to analyze their big data in order to get profits, the paper proposes a case study to show the effective of our method.

Keywors: Channel theory, HowNet, extension model, big data, analysis

1. Introduction

Big data does not only enable users to get information, but also cause difficulty for them to get really required information. This concerns semantics understanding of information. Heterogeneity, scale, timeliness, complexity, and privacy problems with big data impede progress at all phases of the pipeline that can create value from data [1]. The key to making big data initiatives a success lies within making the produced data more digestible and usable in decision making, rather than making it just ‘more,’ resulting in the creation of an environment wherein information is used to generate real impact[2]. Although there are a number of successful customer data analysis tools, but we cannot just analyze data without trying to improve the situation. In order to build analysis tools for companies to get value from big data, we focus not only on information volume, variety and velocity, but also on understanding big data and get value from it.

* Weihua Li. Tel.: +86-20-39322279; fax: +86-20-39322279.
E-mail address: lw@gdut.edu.cn.

doi:10.1016/j.procs.2016.07.118

This paper proposes to combine Channel theory, HowNet structure and extension model to analyze big data. We believe this will more effectively treat big data so that people can get values. As a case study, at the end of this paper, an example of out method is described to show the effects.

2. Channel Theory

Big data, with its increase in the quantity and variety of information, almost inevitably, these discrete resources use different terms to describe similar concepts, or even use identical terms to mean very different things, introducing confusion and error into their use. Barwise and Seligman proposed a very general qualitative theory of information flow (in distributed systems) [3]. It is a mathematic model that aims at establishing the laws that govern the flow of information. Information Flow Theory, which is also called Channel Theory, is a general theory of regularity that applies to the distributed information inherent in both natural world of biological and physical systems and the artificial world of computational systems [3]. It is based on the understanding that information flow results from regularities in a distributed system, and that it is by virtue of regularities among the connections that information of some components of a system carries information of other components.

In channel theory, each component of a distributed system is represented by a classification:

$$\mathbf{A}=<A, \sum_A, \models_A>$$

It consisting of a set A of objects to be classified, called tokens of A, a set $\sum_A$ of objects used to classify the tokens, called the types of A, and a binary relation, $\models_A$ between A and $\sum_A$ that tells one which token are classified as being of which types.

In channel theory, the notion of an infomorphism f: A$\leftrightarrows$B gives a mathematical model of the whole-part relationship between instances of a whole, as modeled by a classification B, and that of a part, as modeled by a classification A.

$$\begin{tikzcd}
\text{typ(A)} \arrow[r, "f^{\wedge}"] \arrow[d, dash, "\models A"] & \text{typ(B)} \arrow[d, dash, "\models B"] \\
\text{tok(A)} & \text{tok(B)} \arrow[l, "f^{\vee}"']
\end{tikzcd}$$

Fig. 1. Infomorphism

An information channel consists of an indexed family C = $\{f_i: \mathbf{A} \leftrightarrows \mathbf{C}\}_{i \in I}$ of infomorphisms with a common codomain C, called the core of the channel.

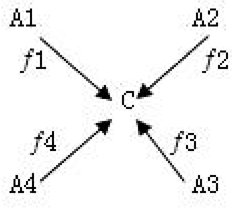

Fig. 2. The basic construct of channel

More details of channel theory can be found in [3]. This mathematical structure effectively captures the local syntax and semantics of a community for the purpose of semantic interoperability [4].

There are some research work and applications of Information-Flow theory in the resent years. One is R. Kent's Information Flow Framework for the IEEE standardization activity of an upper-level ontology [5]. Kent's summary is: Ontology sharing can be successfully founded upon principles of Information Flow. The next one is Y. Kalfoglou and M. Schorlemmer's IF-Map: An Ontology-Mapping Method based on Information-Flow Theory [6]. They formalized the notion of ontology, ontology morphism and ontology mapping and linked them to the formal notions of local logic and logic informorphism stemming from IF theory. The third one is G. Allwein et al's A New Framework for Shannon Information Theory [7]. They synthesize a new theory from Barwise/Seligman theory and Shannon theory so that the qualitative and quantitative analysis uses the same theory structures.

Big data information resources are distributed and highly heterogeneous and dynamic. If we use normal communication means to get the data, we may get the semantic conflicted information which cannot be used. For example, when a user find "cricket moving", an athlete believes that it is a ball. But a biologist may think it an insect. These two persons cannot share this information.

The Channel theory can be used to solve these semantic conflicts. We can use a classification to represent a data resource. We can let infomorphisms to act as semantic interoperability mechanisms. We can build a channel to connect the data resources so that the information can be shared. For semantic interoperating, we suggest HowNet knowledge base be the core of the channel as the basic communication support mechanism.

3. HowNet Structure and Extension Model

HowNet is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents [8]. The philosophy behind HowNet lays ground on its understanding and interpretation of the objective world. The crux is: all matters (physical and metaphysical) are in constant motion and are ever changing in a given time and space. Things evolve from one state to another as recorded in the corresponding change in their attributes [8]. The units for manipulation and description in HowNet are thing (sub-divided into physical and mental), Part, Attribute, Time, Space, Attribute-value and Event.

Every concept as an entry in HowNet has the following description items for each language:

W_X= word or phrase
G_X= the part of speech
E_X= examples
DEF= concept definition

where X can be C or E, C stands for Chinese and E stands for English.

For example, the concept “teach” is defined as:

NO.=043868
G_C=V
E_C=
W_E=teach
G_E=V
E_E=
DEF=teach, education

HowNet is powerful in describing relations between concepts. It describes not only the relations within the same categories, but also describes cross-category relations. The types of relations described by HowNet are mainly as follows [9]:

a. super-ordinate-subordinate
b. synonym
c. antonym
d. converse
e. part-whole (coded with pointer %, e.g. “heart”, “CPU”, etc)
f. attribute-host (coded with pointer &, e.g. “color”, “speed”, etc)
g. material-product (coded with pointer ?, e.g. “cloth”, “flour”, etc)
h. agent-event (coded with pointer *, e.g. “doctor”, “employer”, etc), (may also be “experiencer” or “relevant”, depending on the type of event)
i. patient-event (coded with pointer $, e.g. “patient”, “employee”, etc), (may also be “content” or “possession”, etc. depending on the type of event)
j. instrument-event (coded with pointer *, e.g. “watch”, “computer”, etc)
k. location-event (coded with pointer @, e.g. “bank”, “hospital”, “shop”, etc)
l. time-event (coded with pointer @, e.g. “holiday”, “pregnancy”, etc)
m. value-attribute (coded without pointer, e.g. “blue”, “slow”, etc)
n. entity-value (coded without pointer, e.g. “dwarf”, “fool”, etc)
o. event-role (coded with role-name, e.g. “wail”, “shopping”, “bulge”, etc)
p. concepts related (coded with pointer #, e.g. “cereal”, “coalfield”, etc)

There are three key points that HowNet grasp and reveal knowledge [10]:

•recognized the knowledge is the relational system. The relations have two kinds, one is relation between the concept and the concept, the other one is relation between the attributes and the attributes of the concept.

•take selected sememes and selected semantic relations (a total of 90) between concepts as the concept description units. Use one kind of structural description language to describe each concept quite fine. Sememes are the smallest unit of the most basic meaning, not easy to divide again.

•use the classified system to reveal the axiom.

For example, “customer” is a sememe in HowNet.

NO.=050490
G_C=N
E_C=
W_E=customer

G_E=N
E_E=
DEF=human,*buy, commercial

For example, “low” is a sememe in HowNet.

NO.=020865
G_C=ADJ
E_C=
W_E=low
G_E=ADJ
E_E=
DEF=aValue, height, low

At the same time, “value” is a sememe in HowNet.

NO.=042357
G_C=N
E_C=
W_E=value
G_E=N
E_E=
DEF=attribute, value, &entity

Application of HowNet includes for better understanding and for sense disambiguation.

Extenics and extension methods can describe information. According to current data, it makes standard data processing with basic-element representation system. Basic-elements (including matter-element, affair-element and relation-element) have the following structure:

$$B=\begin{bmatrix} O, & c_1, & v_1 \\ & \vdots & \vdots \\ & c_n, & v_n \end{bmatrix}$$

where O is a object, c_i is the ith characteristic of the object, and v_i is the measure as to the characteristic i.

We can describe the above “customer” by matter-element M_1 as follows:

$$M_1=\begin{bmatrix} \text{customer,} & \text{generic attribute,} & \mathit{human} \\ & \text{behavier,} & \mathit{buy} \\ & \text{possession,} & \mathit{commodity} \end{bmatrix}$$

A low value customer is one type of customers. We can describe this concept by compound element M':

$$M'=\begin{bmatrix} \text{low value customer,} & \text{generic attribute,} & M_1 \\ & \text{current value,} & \mathit{low} \\ & \text{potential value,} & \mathit{low} \end{bmatrix}$$

We describe not only the relationship between the concepts, but also the relationships between attributes and attributes of a concept. Then it is easier to form a knowledge network, supporting data analysis.

For example, use relation-element R_1 to describe the relationships between attributes and attribute of a concept:

$$R_1 = \begin{bmatrix} \text{corelative relation}, & \text{antecedent}, & \text{purchase quantity} \\ & \text{consequent}, & \text{brand loyalty} \\ & \text{degree}, & \textit{close} \end{bmatrix}$$

The rebuilt basic-element base with HowNet information structure forms the knowledge base of extension model. We store this knowledge base to support big data analysis.

4. Combining Channel Theory and HowNet

Information flow theory has been proved that a distributed system can be covered with a channel (and there is a minimum cover). What we care about is how to get the information we need from the established channel, and keeps the semantic consistency with large data sources. From [11] the semantic interoperability of steps, the key starting point is to choose the type, tokens and their classification relationship. Ontology plays an important role. So we believe that big data semantic understanding mechanism needs three kinds of ontology to support.

The first kinds of ontology is a big data sources classification ontology, an information source is equal to in the channel theory the distributional system module showed by a classification Ai = < tok (Ai), typ (Ai), ⊨ Ai >. The information classification ontology must to classify instance tok (Ai) according to the accurate type of typ (Ai), such as "apple" is to classify as electronic equipment, fruit or clothing.

The second kind of ontology is the community ontology. It supports to establish each community context. Because the context of each kind of information source is different, it is very easy to create the semantics conflict. For instance an information source provides “the apple” handset, another information source provides “the apple” fruit. This may elimination conflict through the infomorphism fj between information source ontology and the community ontology.

The third one is channel ontology, can also be referred to as a public ontology. It is the core of sharing information. Every community ontology get agreements through the infomorphism gi with channel ontology. Form ontology sharing. In this way, an IT corporate information say "apple" hiring (sales) will not make a food community feel tempted.

The structure information source classification ontology and the community ontology are easy, because their information example collections are independent control. Then channel ontology establishment is quite complex. It requests in between the participation community's information example to have a natural set “the connection”, namely everybody agreement public inheritance type. After we research, we believe with (HowNet)(18) to take the channel ontology the foundation is feasible.

We proposed the channel ontology is precisely a common knowledge, while the community ontology looks like the special domain the knowledge. We may use sememes in HowNet as far as possible to description foundation data. When sememes in HowNet can not described domain knowledge then to use additional for the description general knowledge the concept. Because sememes in HowNet are originally the careful choice, 2000 sememes each one has one semantic, without the different meanings.

5. Case Study

Our first application of big data semantics shared method is career information sharing. Because in the professional technical institute the competitive power is congenitally deficient, student's employment pressure is high. If we can share professional information in big data that will undoubtedly provide more employment opportunities for graduates. However, these big data must be understood by the professional technical institute's students so that is useful. If someone forwards "Apple company recruiters" information, students must ask what products the "apple company" it engaged in. Is fruit, cell phone, or clothes? Therefore different students may consider whether to apply for a job. This is the importance of big data semantics understanding.

Our extension model describes big data according to the channel theory. It gives token and type together. For example, "Apple company recruiters" will have the following form:

$$M_1=\begin{bmatrix} \text{company}, & \text{name}, & apple \\ & \text{business}, & phone \\ & \text{location}, & City \end{bmatrix}$$

$$A_1=[\ \text{recruit}, \quad \text{acting object}, \quad M_1\ ]$$

When this information from big data is sent to a professional technical institute, the information technology students will consider to applying that job.

If the information is sent like that:

$$M_2=\begin{bmatrix} \text{company}, & \text{name}, & apple \\ & \text{business}, & fruit \\ & \text{location}, & town \end{bmatrix}$$

$$A_2=[\ \text{recruit}, \quad \text{acting object}, \quad M_2\ ]$$

The business or agriculture students may consider to applying that job.

More recruit information will be described by extension model. The information token and type is sent together as channel theory suggest. Every community understands this information according to their ontology especially HowNet. So the professional technical institute can benefit from big data and improve its student employments.

6. Conclusion

This paper proposes to combine Channel theory, HowNet and extension model to analyze big data. The Channel theory gives us a formal support for representing and automating semantic interoperability. The paper proposes HowNet Knowledge base be the core of the channel as the basic communication support mechanism. It is inspired by Information-Flow Theory and related research work. We take advantage of HowNet information structure to rebuild the knowledge base of extension model. This enhances the semantic ability of extension model to treat big data. We have done some experiment work to meet challenges in big data analysis.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant no. 61273306). The authors wish to express thanks to the research group in the School of Computers and the Research Institute of Extenics and Innovation Methods, Guangdong University of Technology.

References

[1] Agrawal, D., Bernstein, P., Bertino, E. et al. Challenges and Opportunities with Big Data. CYBER CENTER TECHNICAL REPORTS, Purdue University (2011)

[2] Forte Wares. Failure to Launch: From Big Data to Big Decisions. http://www.fortewares.com/Administrator/userfiles/Banner/forte-wares-pro-active-reporting_EN.pdf

[3] Barwise, J., Seligman, J. *Information Flow: The logic of Distributed Systems*. Cambridge: Cambridge University Press; 1997.

[4] Kalfoglou, Y., Schorlemmer, M., 2004. “Formal Support for Representing and Automating Semantic Interoperability”, Proc.1st European Semantic Web Symposium (ESWS'04). Heraklion, Crete, Greece, pp 45-61.

[5] Kent, R., 2000. “The information flow foundation for conceptual knowledge organization”, the 6th Int. Conf. International Society for Knowledge Organization. Toronto, Canada, pp 111-117.

[6] Kalfoglou, Y., Schorlemmer, M. IF-Map: an ontology mapping method based on information flow theory. *Journal on Data Semantics* 2003; LNCS 2800:98–127.

[7] Allwein G. T, Moskowitz I. S, Chang L,2004. A New Framework for Shannon Information Theory. NRL Memorandum Rep. NRL/MR/5540-04-8748

[8] DONG Z., DONG Q., HowNet. http://www.keenage.com/

[9] Dong, Z., 1999. “Bigger Context and Better Understanding -- Expectation on Future MT Technology,” Proceedings of the International Conference on Machine Translation & Computer Language Information Processing. Beijing, China, p.17-25.

[10] Dong, Z., Dong, Q., Hao, CL. Theoretical finding of HowNet. *J Chinese Information Processing* 2007; 21:4 3-9.

[11] Schorlemmer, M., Kalfoglou, Y., 2003. “On Semantic Interoperability and the Flow of Information”, Proceedings of the Semantic Integration Workshop, Collocated with the Second International Semantic Web Conference. Sanibel Island, USA, p80-86.