

2129-DOCT

FUZZY ONTOLOGICAL SUPPORT ON DATA MINING

Eduardo de Mattos Pinto Coelho (Universidade Federal de Minas Gerais, Minas Gerais, Brasil) - emattos@pbh.gov.br

Marcello Peixoto Bax (Universidade Federal de Minas Gerais, Minas Gerais, Brasil) - bax@eci.ufmg.br

Wagner Meira Júnior (Universidade Federal de Minas Gerais, Minas Gerais, Brasil) - meira@dcc.ufmg.br

The doctorate research project aims to improve technology support for detecting, analyzing and combating tax evasion in Imposto Sobre Serviços de Qualquer Natureza – ISSQN – a municipal tax, from recovery and processing of large volumes of scattered data in relational databases. The project proposes to exploit the ties of Information Science with aspects of human cognition and recovery and data mining techniques coming from Computer Science, in order to obtain an appropriate representation and retrieval of knowledge through ontologies. To do this, leave the study, comparison, improvement and development of fuzzy ontologies that have the advantage of allowing a simultaneous development and representation of the symbolic and quantitative aspects of a domain. The knowledge representation will be integrated with Tamanduá, a software developed and patented by Universidade Federal de Minas Gerais, which is a tool of data mining, allowing simulations, testing and measurement of results.

Keywords: fuzzy ontologies, data mining, inference mechanisms, causality.

RESUMO

O projeto de pesquisa, em âmbito de doutorado, tem como objetivo o aprimoramento de tecnologia de suporte à detecção, análise e combate a fraudes fiscais no Imposto Sobre Serviços de Qualquer Natureza – ISSQN - a partir de recuperação e processamento de grande volume de dados dispersos em bases de dados relacionais. O projeto propõe a exploração dos vínculos da Ciência da Informação com aspectos da cognição humana e de técnicas de recuperação e de mineração de dados oriundas da Ciência da Computação, com o objetivo de se obter uma adequada representação e recuperação do conhecimento através de ontologias. Para tal, partirá do estudo, comparação, aprimoramento e desenvolvimento de ontologias *fuzzy* que apresentam a vantagem de permitir uma elaboração e representação simultânea de aspectos simbólicos e aspectos quantitativos de uma área de domínio. A representação do conhecimento de determinado domínio possibilitada pelas ontologias *fuzzy* será integrada ao Tamanduá, *software* desenvolvido e patentado pela Universidade Federal de Minas Gerais, que é ferramenta de mineração de dados, permitindo simulações, testes e mensurações dos resultados obtidos. Este projeto busca contribuir com o campo de Ciência da Informação ao promover o levantamento e a discussão do estado da arte em ontologias, em especial ontologias *fuzzy*, ao promover o desenvolvimento original dessas ontologias, sua aplicabilidade e integração com recuperação e mineração de dados. O projeto explora a convergência das habilidades desenvolvidas por três diferentes instituições. A Escola de Ciência da Informação da UFMG, o Departamento de Ciência da Computação da UFMG e a Secretaria Municipal de Finanças da Prefeitura de Belo Horizonte.

PALAVRAS-CHAVE

ontologias fuzzy, mineração de dados, mecanismos de inferência, causalidade

ABSTRACT

The doctorate research project aims to improve technology support for detecting, analyzing and combating tax evasion in Imposto Sobre Serviços de Qualquer Natureza – ISSQN – a municipal tax, from recovery and processing of large volumes of scattered data in relational databases. The project proposes to exploit the ties of Information Science with aspects of human cognition and recovery and data mining techniques coming from Computer Science, in order to obtain an appropriate representation and retrieval of knowledge through ontologies. To do this, leave the study, comparison, improvement and development of fuzzy ontologies that have the advantage of allowing a simultaneous development and representation of the symbolic and quantitative aspects of a domain. The knowledge representation of a given field made possible by fuzzy ontologies will be integrated with Tamanduá, a software developed and patented by Universidade Federal de Minas Gerais, which is a tool of data mining, allowing simulations, testing and measurement of results. This project seeks to contribute to the field of Information Science to promote the survey and discussion of the state of the art in ontology, fuzzy ontologies in particular, to promote the development of original ontologies, their applicability and integration with recovery and data mining. The project explores the convergence of skills developed by three different institutions. The School of Information Science, UFMG, Department of Computer Science, UFMG, and the Department of Finance of the city of Belo Horizonte.

KEY-WORDS

fuzzy ontologies, data mining, inference mechanisms, causality

Escola de Ciência da Informação – UFMG
Doutorado em Ciência da Informação
CIG – Gestão da Informação e do Conhecimento

Ontologias nebulosas no suporte a mineração de dados

Doutorando:

Eduardo de Mattos Pinto Coelho
Gerência de Inteligência Fiscal - Prefeitura Municipal de Belo Horizonte.
Mestre em Engenharia Elétrica pela UFMG.
Doutorando em Ciência da Informação pela UFMG
emattos@pbh.gov.br
Tel.: 55.31.3277-1453

Orientador:

Marcello Peixoto Bax
Professor Associado da Universidade Federal de Minas Gerais, Escola de Ciência da Informação
Doutor em Informática pela Université de Montpellier II
bax@eci.ufmg.br
Tel: 55.31.3409-6103 / 5207

Co-orientador:

Wagner Meira Júnior
Professor adjunto do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação
PhD em Ciência da Computação pela University of Rochester
meira@dcc.ufmg.br
Tel: 55.31.3409.5867

1. Introdução: problemas de pesquisa, objetivos e justificativa.

A pesquisa de doutorado envolve o uso de ontologias no suporte a processos de mineração de dados. Ontologias serão usadas em fase de pré-processamento de dados resultantes das consultas aos bancos de dados, e pós-processamento das regras de associação obtidas no processo de mineração.

Optou-se por trabalhar com ontologias nebulosas para podermos lidar com categorias que apresentam condições de fronteira vagas, difusas, mal definidas, como as que se encontram nos juízos oriundos de avaliações de categorias do direito, de termos da linguagem e de potencialidades inerentes a atos humanos. Ontologias nebulosas apresentam a vantagem de permitir uma elaboração e representação de aspectos conceituais, simbólicos, qualitativos e abstratos de uma área de domínio. Também apresentam a vantagem de lidar, simultaneamente, com aspectos quantitativos, concretos ou numéricos dessa mesma área de domínio. A representação do conhecimento de determinado domínio, possibilitada pelas ontologias nebulosas com suas respectivas regras de inferência, serão integradas à mineração de dados. O sistema resultante será submetido a simulações, testes e mensurações dos resultados obtidos.

Então, especificamente, há de se resolver os seguintes problemas:

1. A modelagem e representação de conhecimento de um domínio através de ontologias nebulosas.
2. A integração dessa ontologia a mecanismos e ferramenta de mineração de dados pré-existente, no caso o Tamanduá, projeto desenvolvido pelo Departamento de Ciência da Computação da UFMG (DCC) na área de mineração de dados. O sistema Tamanduá é uma plataforma patentada de serviços de mineração de dados para apoio à gestão e decisão governamentais.
3. O suporte da ontologia às tarefas de classificação e filtragem dos dados de entrada ao mecanismo de mineração de dados, e dos dados de saída da mineração. As tarefas de classificação e filtragem serão realizadas a partir da análise das entidades e respectivas relações, a fim de se obter os resultados mais relevantes, tanto das consultas, quanto das regras de associação resultantes do processo de mineração.
4. A avaliação da eficiência desse método, dentro dos critérios de cobertura e precisão da informação recuperada, será realizada através da comparação da mineração de dados sem o suporte de ontologias.

O estudo de caso abrange dados de transações econômicas registradas pela Prefeitura Municipal de Belo Horizonte para fins de verificação de incidência, fiscalização e arrecadação de recursos vinculados ao ISSQN – Imposto Sobre Serviços de Qualquer Natureza. A mineração de dados suportada por ontologias nebulosas deverá ser capaz de detectar práticas eventuais e sistêmicas de evasão, sonegação, elisão e fraude fiscal. Com base nessas informações, as ações fiscais são direcionadas, facilitando e otimizando o trabalho dos agentes fiscais em suas atuações externas.

A presente pesquisa apresenta relevância para os trabalhos de detecção e combate a fraudes exercidas por várias entidades do Poder Público, dentre elas, as auditorias tributárias municipais, estaduais e federais, previdência pública, Polícia Federal e Banco Central.

2. Fundamentação teórico-metodológica.

No âmbito da pesquisa citada, os trabalhos em andamento envolvem a construção de duas ontologias nebulosas e respectivos mecanismos de inferência. Ver Figura 1 abaixo.

A primeira ontologia será construída a partir de base de dados (bancos de dados relacionais). Na sua construção, poderão ser utilizadas ferramentas de engenharia reversa para a geração semi-automática de ontologias que será editada e consolidada usando-se a ferramenta Prótegé. Ela será utilizada em nível de pré-processamento, à etapa de mineração de dados, do resultado das consultas às bases de dados relacionais. A ferramenta de extração semi-automática de ontologias a ser utilizada será avaliada, dentre algumas disponíveis na Web e de acesso gratuito.

A segunda ontologia será construída a partir de base de dados semi-estruturados ou não estruturados, constituída de documentos de análise e procedimentos de inteligência fiscal e de conhecimentos de contexto/fundo, focada na semântica das regras de associações a serem destacadas pela ferramenta de mineração de dados Tamanduá. Ela será utilizada em nível de pós-processamento, à etapa de mineração de dados, do resultado dessa mineração, constituída pelas regras de associação descobertas pelo mecanismo de mineração. Neste aspecto, há as importantes referências que tratam das relações parte-todo (mereologia e mereotopologia), relações de causalidade, Min-Max, tipo Cambridge etc., que serão úteis na construção dessa ontologia que irá auxiliar-nos na avaliação semântica das associações resultantes da mineração.

As ontologias e respectivos mecanismos de inferências serão construídos com base na teoria de sistemas nebulosos e o referencial metodológico de desenvolvimento, em princípio,

será a metodologia Methontology. A Methontology enfatiza o compartilhamento, reutilização e desenvolvimento de protótipos em ontologias e prevê a utilização da ferramenta Protégé.

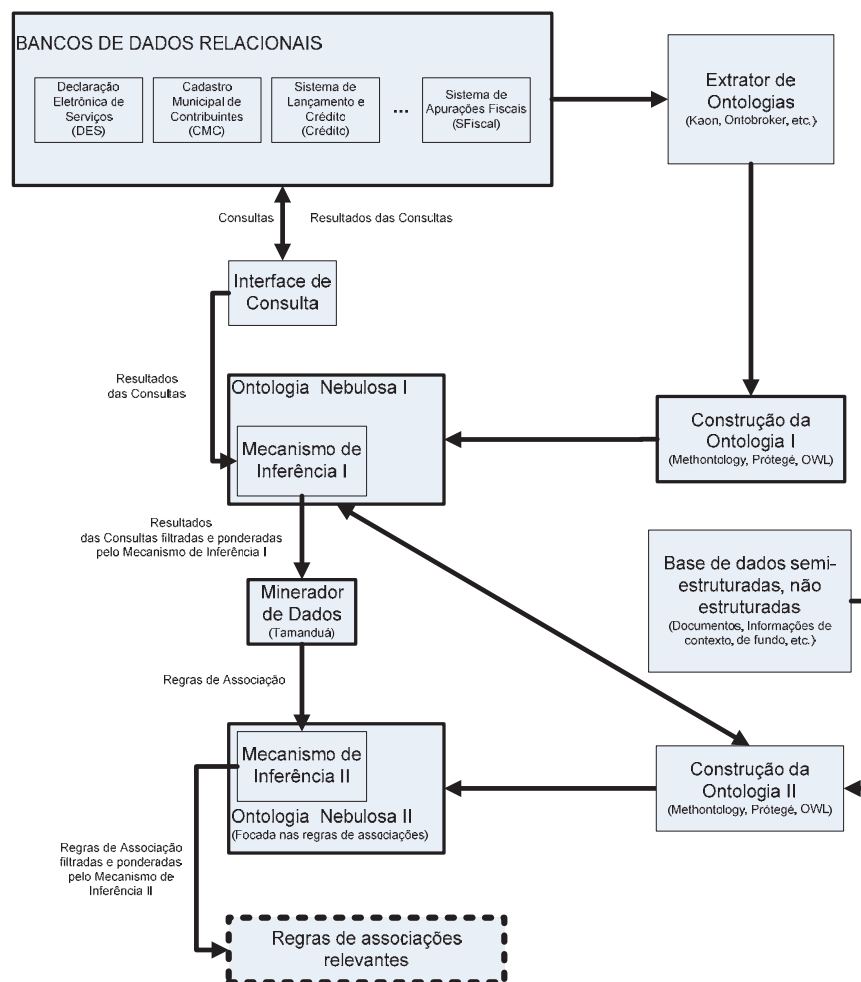


Figura 1. Visão panorâmica das tarefas envolvidas na pesquisa.

3. Contribuições esperadas, informando no âmbito social.

Esse trabalho de pesquisa reveste-se de ampla relevância social por buscar solução que se constitua instrumento de promoção da justiça e equidade tributária através do aperfeiçoamento do sistema fiscal, da prevenção e do combate à fraude e à evasão fiscal. Busca o incremento dos recursos, a serem disponibilizados para a sociedade, através da devida inclusão na esfera da incidência tributária de fatos geradores de impostos que, mediante dissimulação, sonegação ou omissão fogem do alcance tributário. Essa fuga beneficia indevidamente agentes econômicos com o aumento de sua disponibilidade econômica, afetando a livre concorrência no mercado, favorecendo, ainda mais, a terrível

concentração de renda que persiste em nosso país. De outro lado, sua devida inclusão promove a equidade, a justiça fiscal e o aprimoramento na distribuição da renda.

4. Direções de pesquisa.

O desenvolvimento das ontologias baseia-se em diferentes abordagens que se fundam em distintos pressupostos lógicos e filosóficos. Podem-se destacar ao menos três grandes vias que orientam seu desenvolvimento. Primeiro, as ontologias formais de linha fenomenológica baseadas nos trabalhos filosóficos de Brentano e Husserl, representados, em especial, no campo da Ciência da Informação, por Barry Smith e Nicola Guarino. Segundo, as ontologias formalizadas baseadas em técnicas de representação desenvolvidas no âmbito da Inteligência Artificial, representadas, em especial, no campo da Ciência da Informação, por John Sowa. Terceiro, as ontologias formalizadas baseadas nos desenvolvimentos da Primeira Filosofia Analítica, de índole mais logicista e formalista, e da Segunda Filosofia Analítica, de índole mais pragmática e contextualista, representadas em trabalhos esparsos. Essas três vias propõem visões distintas que se refletem no desenvolvimento de metodologias, métodos e abordagens de problemas práticos. A questão da vaguidade, por exemplo, relacionada às ontologias nebulosas, recebe tratamentos diferentes, conforme a via adotada.

O discernimento dessas vias, com a elucidação de suas vantagens e desvantagens, adequabilidade e resultados alcançados é fundamental para o entendimento das soluções até então adotadas e para a proposição de novas soluções e, mesmo, novas vias de desenvolvimento de soluções.

5. Bibliografia.

ADA, Chan Man Kuok, e FU, Man Hon Wong. “Mining Fuzzy Association Rules“, 1998. In: *Databases*. ACM, SIGMOD Volume 27, Issue 1, March, 1998, pp. 41-46.

ARTALE, A., Franconi, E., Guarino, N., and Pazzi, L. 1996. “Part-Whole Relations”. In *Object-Centered Systems: an Overview. Data and Knowledge Engineering*, 20(3), pp. 347-383.

BAZIZ, Mustapha; BOUGHANEN, Mohand; PRADE, Henri e PASI, Gabriella. “A Fuzzy Logic Approach to Information Retrieval Using na Ontology-based Representation of Documents”. In: SANCHEZ, Elie (ed.) *Fuzzy Logic and the Semantic Web*, 2005, pp. 363-378.

BRANCO, Antonio Carlos Saraiva. “Geração de Fuzzy Queries para Mineração de Dados”. Tese de Doutorado, COPPE, Rio de Janeiro, 2004, 72 pp.

CALEGARI, S., CIUCCI, D.: “Fuzzy Ontology and Fuzzy-OWL in the KAON Project”. In: *FUZZY IEEE 2007. IEEE International Conference on Fuzzy Systems* (2007) In printing.

CASANOVA Anderson Araújo, LABIDI, Sofiane. “Algoritmo da Confiança Inversa para Mineração de Dados Baseado em Técnicas de Regras de Associação e Lógica Nebulosa”. XXV Congresso da Sociedade Brasileira de Computação, V ENIA, 22 a 29 de julho, UNISINOS, São Leopoldo;RS, pp. 1106-1109.

COELHO, Eduardo de Mattos Pinto Coelho. “Abordagem neuro-nebulosa para diagnóstico automatizado de câncer de colo uterino”. Dissertação de Mestrado, Escola de Engenharia, UFMG, 1997.

CROSS, Valerie V. “Similarity or Inference for Assessing Relevance in Information Retrieval”. *IEEE International Fuzzy Systems Conference*. 2001, pp. 1287-1290.

DCC/UFMG, “Manual do Tamanduá” (2006). Versão 1.1 - 23 de Agosto de 2006. Departamento de Ciência da Computação. Disponível na Web em <http://tamandua.speed.dcc.ufmg.br>, acesso em setembro-2006.

ESCOVAR, Eduardo L. G.; YAGUINUMA, Cristiane A. e BIAJIZ, Mauro. “Using Fuzzy Ontologies to Extend Semantically Similar Data Mining.” Department of Computer Science – Federal University of São Carlos (UFSCar), pp. 16-30.

KRONBAUER, Arthur Henrique; FERREIRA, Leandro da C. e JÚNIOR, Reinaldo N., “Mineração de Dados Utilizando Lógica Fuzzy”. In: *Científico*, Ano V, Volume II, 2005.

LAM, Toby H.W. “Fuzzy Ontology Map - A Fuzzy Extension of the Hard-Constraint Ontology”. *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI 2006 Main Conference Proceedings, WI'06, 2006.

MARINICA, Claudia; GUILLET, Fabrice; BRIAND, Henri. "Post-Processing of Discovered Association Rules Using Ontologies," *IEEE International Conference on Data Mining Workshops*, 2008, pp.126-133.

PAN, Jeff Z.; STAMOU, Giorgos; STOILLOS, Giorgio; TAYLOR, Stuart e THOMAS, Edward. “Scalable Querying Services over Fuzzy Ontologies”. *ACM, WWW 2008*, April

21-25, Beijing, China, pp. 575-584.

PENZO, Wilma. "Rewriting Rules to Permeate Complex Similarity and Fuzzy Queries within a Relational Database System". *IEEE Transactions On Knowledge and Data Engineering*, Vol. 17, n. 2, February, 2005, pp. 255-270.

PEREIRA, R.; RICARTE, I. L. M.; GOMIDE, F. A. C. "Fuzzy Relational Ontological Models in Information Search Systems". In: Elie Sanchez. (Org.). *Fuzzy Logic and the Semantic Web*. Amsterdam: Elsevier, 2006, pp. 395-412.

PEREIRA, R.; RICARTE, I. L. M.; GOMIDE, F. A. C. "Relational ontology in information retrieval systems". In: *Eleventh International Fuzzy Systems Association World Congress*, 2005, Beijing. *Fuzzy Logic, Soft Computing and Computational Intelligence*. New York: Springer, 2005. v. I. pp. 509-514.

PRIBBENOW, S. "Meronymic relationships: from classical mereology to complex part-whole relations" in GREEN, R.; BEAN, C. A.; MYAENG, S. H. *The semantic of relationship: a interdisciplinary perspective*. Dordrecht: Kluwer, 2002, pp. 35-50.

SCHULZ, Stefan. "Mereotopological Reasoning about Parts and (W)Holes in Bio-Ontologist". *ACM, FOZS '01*, October 17-19, 2001, 2001, pp. 210-221. Disponível em: <<http://www.imbi.uni-freiburg.de/medinf/Nschulz.htm>>.

SMITH, Barry. "Mereotopology: a theory of parts and boundaries". *Data & Knowledge Engineering*, v. 20, n. 3, nov., 1996, pp. 287-303.

SMITH, Michael H.; RUBIN, Stuart e TRAJKOVIC, Ljiljana. "Fuzzy Data Mining for Querying and Retrieval of Research Archival Information". *IEEE*, 1998, pp. 140-145.

TAKAGI, T. e Kawase, K. "A trial for data retrieval using conceptual fuzzy sets". *IEEE Transactions on Fuzzy Systems*, 9(4), 2001, pp. 497-505.

UNCU, Ozge e TURKSEN, I.B.A "New Two-step Fuzzy Inference Approach Based on Takagi-Sugeno Inference Using Discrete Type 2 Fuzzy Sets". *IEEE*, 2003, pp. 32-37.

YAO, Min; WANG, Weihong; ZHAO, X. "A Kind of Method for Mining Classification Rules Based on Fuzzy Sets". *ICSP'02 Proceedings*, IEEE, 2002, pp. 1263-1266.

ZADEH, L. A. "Making computers think like people". In: *IEEE Spectrum*, August, 1984, pp. 26-32.

ZADEH, L. A. "Fuzzy Sets". In: *Information Control*, Volume 8, 1965, pp. 338-353.