Návrh výskumného projektu: Porovnávanie informácií s využitím znalostí na pozadí

Martin Kováčik

Slovenská technická univerzita
Fakulta informatiky a informačných technológií
Ilkovičova 3, 842 16 Bratislava
mato.kovacik@gmail.com

Abstrakt. Množstvo informácií, ktoré je potrebné spracovávať neustále narastá. Informácie sú prezentované v neštruktúrovanej forme, z ktorej nie je jednoduche strojovo určiť význam informácie a jej prepojenie. Jednou z pomocných úloh pri určovaní významu a vyhľadávaní v informáciách je klastrovanie. Klastrovanie predpokladá, že informácie, ktoré sú v združené v klastroch sú prepojené. Jedným z možných spôsobov prepojenia informácií je ich podobnosť. Cieľom tohto návrhu výskumného projektu je rozobrať možné spôsoby na porovnávanie informácií a vhodnosť ich aplikácie v závislosti od použitých dát. Predpokladom je využitie už nadobudnutých znalostí. Tieto znalosti je možné neskôr rozširovať a meniť na základe podnetu od používateľa.

Kľúčové slová: semantic metrics, ontology matching, semantic matching, context vector, clustering

Úvod

Internet nám sprístupňuje obrovské množstvá informácií. Problémom pre prácu s informáciami je ich duplicita a obmedzené významové prepojenia. Významové prepojenie je komplikované distribuovaným ukladaním dokumentov a heterogénnou prezentáciou informácií.

Jednou z ciest na zvládnutie veľkého množstva informácií je ich zhlukovanie do klastrov. Toto umožňuje na základe stanovených kritérií vytvárať zhluky príbuzných dokumentov. Jedným z potrebných nástrojov pri klastrovaní je porovnávanie dokumentov. Porovnávanie dokumentov avšak nie je obmedzené len na klastrovanie informácií, ale je možné ho využiť aj pri iných činnostiach.

Zámerom tohto návrhu výskumného projektu je porovnávanie informácií na základe dodatočných znalostí. Informácie dostupné v distribuovanom prostredí častokrát nemajú potrebnú formu vhodnú na porovnávanie – predpísanú štruktúru,

alebo formu. Prítomnosť znalostí na pozadí, umožňuje ich využitie na dodatočné pridanie skrytých informácií a vytvára tak štruktúry vhodnejšie na porovnávanie.

Prehľad problematiky

Problém integrácie informácie z rôznych zdrojov je pri aktuálnom stave veľmi žiadúci. Táto téma je rozoberaná z pohľadov rôznych vedeckých odborov – z pohľadu umelej inteligencie a znalostného inžinierstva.

Existuje viacero spôsobov, ktoré umožňujú porovnávanie informácií. Jeden prístup je určovanie štatistickej podobnosti dokumentu. Štatistické metódy sa intenzívne využívajú na ochranu pred spamom.

Iný prístup je s využitím ontológií. Informácia (ďalej dokument) je spracovaná pomocou ontológie a sú získané príslušné koncepty. Následne sú porovnávané ontológie, s ktorými sa mapovali porovnávané dokumenty. Prístupy sú viacere:

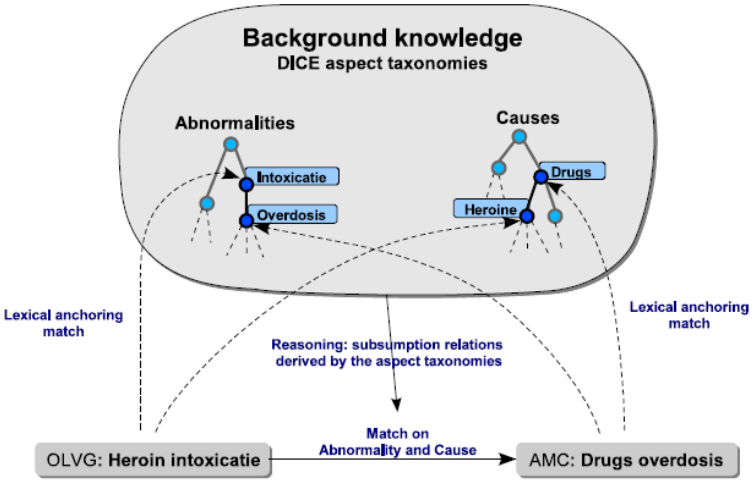
- terminologické – tieto metódy porovnávajú lexikálnu korešpondenciu medzi triedami v oboch porovnávaných ontológiách
- porovnávanie na základe inštancií – vyhľadávajú podobné inštancie
- štrukturálne – využívajú štruktúru ontológií
- semantické – využívajú indukciu na odvodzovanie vzťahov

Podľa [1] prevažujú prístupy využívajúce kombinácie terminologických a štruktúrovaných metód, kde lexikálny prekryv ontológií slúži na iniciálne mapovanie. Predpokladom pre využitie týchto prístupov je dostatočný prekryv ontológií. Ďalšia požiadavka je aby ontológie porovnávaných dokumentov mali dostatočnú štruktúru.

[1] navrhuje systém, ktorý využíva znalosti na pozadí. Úlohou týchto informácií je poskytnutie mapovania medzi zdrojovým a cieľovým dokumentom - zjednocovanie použitých konceptov. Príklad aplikácie tohto prístupu v praxi je znázornený na obr. 1. Na tomto obrázku je znázornený jeden z prístupov porovnania / mapovania jednotlivých informácií s využitím znalosti na pozadí.

Iným prístupom je rozvíjanie konceptov s využitím ontológie na pozadí a následné porovnanie výslednej štruktúry. Tento prístup porovnáva štruktúru a určuje príbuznosť viacerých typov dokumentov. Takýto prístup je prezentovaný v [2]. Uvedený prístup rozvíja postupne porovnávane koncepty do vektora znakov. Následne je ohodnotený vektor znakov pre zdrojový a cieľový kontext porovnaný.

Pri postupnom rozvoji kontextov je potrebné ohraničiť rozvoj, aby nevznikali cykly. Ohodnotenie jednotlivých znakov vo vektore je dané prítomnosťou inštancií v rozvoji. Príklad na obr. 2 prezentuje spôsob rozvoja jednotlivých konceptov. Na obr. 3 sú výsledné vektory, ktoré je následne možné porovnávať.



Obr. 1: Príklad mapovania dvoch medicinskych konceptov pomocou znalosti na pozadí

$$\begin{aligned}
\text{Book} &\doteq \text{Document} \sqcap \geq_1 \text{hasTitle} \sqcap \geq_1 \text{hasYear} \\
&\quad \sqcap \geq_1 \text{hasPublisher} \sqcap \geq_1 \text{humanCreator.Author} \\
\text{Phdthesis} &\doteq \text{Document} \sqcap \geq_1 \text{hasAuthor} \sqcap \geq_1 \text{hasTitle} \sqcap \\
&\quad \geq_1 \text{hasSchool} \sqcap \geq_1 \text{hasYear} \\
\text{Mastersthesis} &\doteq \text{Document} \sqcap \geq_1 \text{hasAuthor} \sqcap \geq_1 \text{hasTitle} \sqcap \\
&\quad \geq_1 \text{hasSchool} \sqcap \geq_1 \text{hasYear}
\end{aligned}$$

$$\text{Author} \doteq \text{Human} \sqcap \geq_2 \text{hasPublication.Document}$$
$$\text{Document} \sqsubseteq \top \quad \text{Human} \sqsubseteq \top$$

$${}^0\mathfrak{C}_1^{\text{Book}} = \left\{ \begin{array}{l} x : \text{Document} \sqcap \geq_1 \text{hasTitle} \sqcap \geq_1 \text{hasYear} \sqcap \\ \quad \geq_1 \text{hasPublisher} \sqcap \geq_1 \text{humanCreator.Author} \end{array} \right\}$$

$${}^1\mathfrak{C}_1^{\text{Book}} = \left\{ \begin{array}{l} x : \text{Document} \sqcap \geq_1 \text{hasTitle} \sqcap \geq_1 \text{hasYear} \sqcap \\ \quad \geq_1 \text{hasPublisher} \sqcap \\ \quad \geq_1 \text{humanCreator.}(\text{Human} \sqcap \geq_2 \text{hasPublication.Document}) \end{array} \right\}$$

$${}^2\mathfrak{C}_1^{\text{Book}} = \left\{ \begin{array}{l} x : \text{Document},\ x : \geq_1 \text{hasTitle},\ x : \geq_1 \text{hasYear}, \\ x : \geq_1 \text{hasPublisher}, \\ x : \geq_1 \text{humanCreator.}(\text{Human} \sqcap \geq_2 \text{hasPublication.Document}) \end{array} \right\}$$

$${}^3\mathfrak{C}_1^{\text{Book}} = \left\{ \begin{array}{l} x : \text{Document},\ \langle x, y_0 \rangle : \text{hasTitle},\ \langle x, y_1 \rangle : \text{hasYear}, \\ \langle x, y_2 \rangle : \text{hasPublisher},\ \langle x, y_4 \rangle : \text{humanCreator}, \\ y_4 : \text{Human} \sqcap \geq_2 \text{hasPublication.Document} \end{array} \right\}$$

$${}^4\mathfrak{C}_1^{\text{Book}} = \left\{ \begin{array}{l} x : \text{Document},\ \langle x, y_0 \rangle : \text{hasTitle},\ \langle x, y_1 \rangle : \text{hasYear}, \\ \langle x, y_2 \rangle : \text{hasPublisher},\ \langle x, y_4 \rangle : \text{humanCreator}, \\ y_4 : \text{Human},\ \langle y_4, z_0 \rangle : \text{hasPublication.Document} \\ \langle y_4, z_1 \rangle : \text{hasPublication.Document} \end{array} \right\}$$

$${}^5\mathfrak{C}_1^{\text{Book}} = \left\{ \begin{array}{l} x : \text{Document},\ \langle x, y_0 \rangle : \text{hasTitle},\ \langle x, y_1 \rangle : \text{hasYear}, \\ \langle x, y_2 \rangle : \text{hasPublisher},\ \langle x, y_4 \rangle : \text{humanCreator}, \\ y_4 : \text{Human},\ \langle y_4, z_0 \rangle : \text{hasPublication},\ z_0 : \text{Document} \\ \langle y_4, z_1 \rangle : \text{hasPublication},\ z_1 : \text{Document},\ x : \top \end{array} \right\}$$

$${}^n\mathfrak{C}_1^{\text{Phdthesis}} = \begin{array}{l} x : \text{Document},\ \langle x, y_0 \rangle : \text{hasAuthor},\ \langle x, y_1 \rangle : \text{hasTitle}, \\ \langle x, y_2 \rangle : \text{hasSchool},\ \langle x, y_3 \rangle : \text{hasYear},\ x : \top \end{array}$$

$${}^n\mathfrak{C}_1^{\text{Mastersthesis}} = \begin{array}{l} x : \text{Document},\ \langle x, y_0 \rangle : \text{hasAuthor},\ \langle x, y_1 \rangle : \text{hasTitle}, \\ \langle x, y_2 \rangle : \text{hasSchool},\ \langle x, y_3 \rangle : \text{hasYear},\ x : \top \end{array}$$

Obr. 2: Príklad rozvoja konceptov

	$\mathfrak{C}_1^{\mathsf{Book}}$	$\mathfrak{C}_1^{\mathsf{Phdthesis}}$	$\mathfrak{C}_1^{\mathsf{Mastersthesis}}$
$\top$ (top)	1	1	1
Document	3	1	1
Human	1	0	0
hasAuthor	0	1	1
hasPublisher	1	0	0
hasPublication	2	0	0
hasTitle	1	1	1
humanCreator	1	0	0
hasSchool	0	1	1
hasYear	1	1	1

Obr. 3 : Výsledný vektor

Výsledkom porovnania v tomto prípade je vzdialenosť konceptov v priestore určenom týmito vektormi (priklad obr. 3).

$$\Delta(\mathsf{Book}, \mathsf{Phdthesis}) = -\log(\mathtt{sim}(\mathsf{Book}, \mathsf{Phdthesis})) \approx 2.101$$
$$\Delta(\mathsf{Book}, \mathsf{Mastersthesis}) = -\log(\mathtt{sim}(\mathsf{Book}, \mathsf{Mastersthesis})) \approx 2.101$$
$$\Delta(\mathsf{Phdthesis}, \mathsf{Mastersthesis}) = -\log(\mathtt{sim}(\mathsf{Phdthesis}, \mathsf{Mastersthesis})) \approx 0$$

Obr. 4 : Vzdialenosť konceptov

Jedným z projektov, ktorý sa zaoberá určovaním podobnosti je Wordnet::Similarity [3]. Tento projekt využíva na určenie podobnosti dokumentov Wordnet [4] ako znalosti na pozadí. Tento projekt slúži ako východiskový bod pre riešenia, ktoré porovnávajú dokumenty na základe výskytu špecifických slov. Základom tejto metódy je tzv. kontextový vektor. Táto metóda určuje priestor v ktorom sú umiestnené jednotlivé slová. Kontext je v tomto prípade rozložený do slov a ich umiestnenie v priestore určuje čiastkové kontextové vektory, ktoré spriemerovaním určujú kontextový vektor. Následne euklidovská vzdialenosť určuje podobnosť kontextov. Táto metóda je bližšie popísaná v [5].

Obmedzenie uvedenej metódy je dané tým, že WordNet databáza je definovaná len pre anglický jazyk.

Obmedzenia a problémy pri porovnávaní informácií

Problematickým miestom pri porovnávaní informácií je mapovanie informácie do konceptov. Predpokladom pre aplikáciu tohto výskumného projektu je porovnávanie textov písaných ľuďmi. Je potrebné zohľadňovať nejednoznačnosť prirodzeného jazyka. Toto je problematické zvlášť v prípade slovenského jazyka.

Záver a sumarizácia cieľov výskumného projektu

Uvádzané prístupy prezentujú rôzne pohľady na tematiku porovnávanie informácií. Cieľom výskumného problému by mala byť podrobnejšia analýza uvádzaných postupov, taktiež analýza iných prístupov, ktoré využívajú dodatočné informácie na porovnávanie informácií. Prvý prístup predpokladá využitie rôznych ontológií a vytvorenie mapovania medzi nimi. Druhý prístup predpokladá využitie rovnakej ontológie pre oba koncepty. Tretí prístup využíva koncepciu príbuzností slov namiesto ich definovaného významu pomocou ontológie.

Cieľom výskumného projektu je vytvorenie rámcového systému na porovnávanie dokumentov. Predpokladá sa aplikácia v prostredí Cluster Navigator [6]. Prehľad problematiky uvádza viacero metód, avšak ich aplikácia nie je univerzálna. Výskumným zámerom je vytvorenie hybridného prístupu, ktorý by umožňoval využitie uvedených prístupov na ohodnocovanie podobnosti dokumentov za účelom ich klastrovania.

Súčasťou výskumu je aj analýza a implementácia funkcionality potrebnej na aktualizáciu informácie na pozadí a jej spracovanie.

Použitá literatúra

1. Aleksovski, Z., Klein, M., Kate, W., Harmelen, F.: Matching Unstructured Vocabularies Using a Background Ontology. *Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management ((EKAW)'06)*
2. Bo Hu, Kalfoglou, Y., Alani, H., Dupplaw, D., Lewis, P., Shadbolt, N.: Semantic Metrics, *Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management ((EKAW)'06)*
3. Pedersen, T.: WordNet::Similarity, URL: http://www.d.umn.edu/~tpederse/similarity.html, stiahnuté 4.12.2006
4. University of Princeton: Wordnet, URL: http://wordnet.princeton.edu/
5. Patwardhan, S., Pedersen, T.: Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts, *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together, April 4, 2006, Trento, Italy*
6. Frivolt, G.: Cluster Navigator: Identification of Graph Clusters, *Proceedings from Research Project Workshop Tools NAZOU, September 29, 2006, Nízke Tatry*