# Master Thesis Review / Posudek oponenta diplomové práce

**Jan Rybák:** Visualization of Cost Data

## English version

The master thesis describes on 85 pages of text the motivation, theoretical background (methods, data) and technical realization for a system visualizing interesting data on financial aspects of real world entities, as extracted from natural language text corpora.

On the positive side, the author has clearly mastered a wide variety of natural language processing methods as well as core information visualization approaches. The list of referenced works is impressive (52 items) and includes current research in the area. The practical implementation is well thought out both architecturally and technologically, the author uses suitable leading edge solutions (e.g. the GATE Framework by Sheffield University for the text processing pipeline and HTML5/JavaScript for the visualization application's web frontend). The final result is a working demonstration application which enables the users to explore unexpected relations between the data.

On the negative side, the text of the thesis is imbalanced and somewhat weak. Chapter 2 (State of the Art) contains many shallow explanations and in general does not provide sufficient background to methods used in the thesis. In Chapter 3 (Realization), the overall picture is missing and the reader gets lost in a long sequence of detailed descriptions of method details, practical implementation issues and experiences gained; the former aspect should have been covered in Chapter 2 and better examples should have been used (explain actual data in the corpus, final processed RDF triples, etc.). Evaluation in Chapter 4 is rather short and does not touch the validation question whether the implemented method really provides better insight into the finance-related data. Finally, the part on Information Visualization (pp. 63+) has an interesting introduction but lacks wider treatment of visualization techniques and fails to discuss the treemap weaknesses. Many particular issues could also be raised, see the questions at the end of the review for some of them.

The formal aspects of the text could have been handled better as well: sources of images are not cited, terms and abbreviations are used before/without proper definition in many places (e.g. "gazetteer", "NER"), labels in Figure 6.1 on p.67 make the description hard to understand (sic). Typography looks nice at first sight but suffers in many places (unexpected page breaks, uneven line and heading spacing, missing page numbers, etc.). The usage of English is mostly commendable but tends to weaken in later parts of the text.

In a summary, there is a lot of good work behind the presented thesis both of from the theoretical and practical standpoints but the text does not make justice to the final result.

## Česká verze

Diplomová práce popisuje na 85 stranách textu motivaci, teoretické základy (metody, datové struktury) a technickou realizaci systému, který poskytuje vizualizaci zajímavých dat o finančních aspektech objektů reálného světa získaných analýzou textových korpusů.

SOUHLASÍ
S ORIGINÁLEM

Práce má mnoho kladných rysů. Prokazuje autorovu schopnost dobře zvládnout širokou škálu metod zpracování přirozeného jazyka i základní metody vizualizace informací. Seznam literatury je nadprůměrně kvalitní (52 položek včetně mnoha výsledků aktuálního výzkumu v oblasti). Taktéž realizace softwarových nástrojů je kvalitní po architektonické i technologické stránce; autor používá vhodné nástroje a moderní technologie vč. například rámce GATE z University of Sheffield pro zpracování textu či HTML5/JavaScript pro webové rozhraní výsledné aplikace. Ta má formu robustního výzkumného prototypu, který poskytuje uživatelům možnost zkoumat nečekané souvislosti v analyzovaných datech.

Na druhou stranu vlastní text práce - psaný v anglickém jazyce - je nevyvážený a obsahově slabší. Kapitola 2 (State of the Art) trpí množstvím spíše povrchních popisů a jako celek nedává dostatečný teoretický základ pro další části práce. Kapitola 3 (Realization) postrádá nadhled, takže čtenář se snadno ztratí v dosti dlouhé posloupnosti popisů konkrétních metod a jejich detailů, praktických obtíží s realizací a získaných zkušeností. Zejména první z těchto bodů měl být řešen v předchozí kapitole; taktéž příklady měly být zvoleny vhodněji tak, aby ilustrovaly např. data v korpusu či výsledné RDF trojice. Vyhodnocení metod v kap. 4 je relativně mělké a pomíjí podstatnou validační otázku, zda výsledná metoda pomáhá uživatelům získat lepší vhled do dat zaměřených na finanční hodnotu věcí. Oddíl o vizualizaci informací (od str. 63 dále) obsahuje obecně velmi zajímavý úvod ale postrádám v něm základní přehled technik pro vizualizaci dat a diskusi nedostatků treemap zobrazení. Bylo by možno dále diskutovat množství konkrétních problematických bodů, některé viz otázky na konci posudku.

Také formální stránce práce měla být věnována větší péče, např. nejsou citovány zdroje obrázků, pojmy a zkratky jsou mnohde používány bez předchozí resp. podrobné definice (např. „gazetteer", „NER"), chyby v popiscích obr. 6.1 na str. 67 komplikují jeho pochopení (sic), atd. Po typografické stránce vypadá práce na první pohled dobře, ale množství drobných problémů svědčí o nedostatku pečlivosti (zbytečná zalomení stránky, rozdíly v řádkování vč. nadpisů, chybějící čísla stránek, atd.). Úroveň angličtiny je povětšinou velmi dobrá, ale ke konci práce roste počet překlepů a stylistických chyb.

Celkově je tedy možno konstatovat, že předložená diplomová práce prokazuje značnou, až nadprůměrnou odbornou erudici autora ale její textová část výsledný dojem poněkud snižuje. Navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě.

Questions to the defense / Otázky k obhajobě:

1. Please explain the use of context information for NER disambiguation using DBpedia Spotlight - it is only briefly introduced on p.43 but not explained subsequently. / Vysvětlete využití kontextové informace z DBpedia Spotlight pro disambiguaci NER - popis na str.43 a následujících je příliš stručný.
2. What is the concrete data set to be visualized and its constraints for treemap algorithms (sections 6.2 and 7.5)? For instance, how are highly imbalanced trees handled if I do not want the treemap to contain very small rectangle „crumbs"? / Jaká data se v práci zobrazují pomocí treemap algoritmů (oddíly 6.2 a 7.5) a jaká jsou jejich omezení? Jak byste řešil problém silně nevyvážených stromů či množin dat tak, aby zobrazení neobsahovalo příliš drobné obdélníky?

V Plzni 7.6.2013

doc. Ing. Přemysl Brada, MSc. Ph.D.

SOUHLASÍ S ORIGINÁLEM

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky