Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky

Extrakce informací z webových stránek pomocí extrakčních ontologií

Autoreferát k doktorské disertační práci

Obor	:	Informatika
Doktorand	:	Ing. Martin Labský
Školitel	:	Prof. Ing. Petr Berka, CSc.
Oponenti	:	Prof. Václav Snášel
		RNDr. Petr Strossa, CSc.
		Prof. RNDr. Peter Vojtáš, DrSc.

Praha, květen 2009

Extrakce informací z webových stránek pomocí extrakčních ontologií

Doktorská disertační práce byla vypracována v rámci doktorského studia oboru Informatika na Fakultě informatiky a statistiky Vysoké školy ekonomické v Praze.

Obhajoba disertační práce se koná 11. června 2009 před komisí pro obhajoby doktorských prací v oboru Informatika na VŠE v Praze, nám. W. Churchilla 4, 130 67 Praha 3.

Obsah autoreferátu

Abstrakt a klíčová slova

S rozvojem Internetu a růstem množství textových dokumentů vzrostla během posledního desetiletí poptávka po nástrojích pro automatickou extrakci informací (IE – information extraction). Úkolem IE je nalézt v analyzovaných dokumentech údaje předem specifikovaného sémantického typu a tyto extrahovat pro potřeby dalších aplikací.

Analyzovanými dokumenty mohou přitom být webové stránky, e-maily, dokumenty uvnitř firemních informačních systémů, ale i běžné textové zdroje jakými jsou noviny a časopisy. Aplikace, které nejčastěji využívají výsledky IE, zahrnují tradiční textové vyhledávače, které lze pomocí IE rozšířit o tzv. strukturované vyhledávání, dále systémy pro automatické zodpovídání otázek, strojový překlad nebo např. aplikace pro podporu posuzování kvality webových stránek.

Tato práce se zabývá vývojem metod a nástrojů pro IE, které jsou obzvláště vhodné pro extrakci ze semi-strukturovaných dokumentů, jakými jsou webové stránky, a pro situace kdy existuje málo trénovacích dat. Hlavním přínosem této práce je navržený přístup *rozšířených extrakčních ontologií*, který v sobě kombinuje využití extrakčních znalostí tří typů: (1) expertem zadané extrakční znalosti, (2) znalosti naučené z trénovacích dat a (3) znalosti indukované na základě opakující se formátovací struktury, která se často ve webových stránkách nachází.

Naše hypotéza je, že současné využití všech tří typů znalostí extrakčním algoritmem může přispět k celkové přesnosti a robustnosti extrakce a k rychlejšímu vývoji extrakční aplikace. Motivací pro tuto práci byl malý počet dosud popsaných metod pro IE, které by využívaly všechny zmíněné typy extrakční znalosti.

Jako první je v tomto textu popsán statistický trénovaný přístup k IE založený na skrytých markovských modelech, který je dále integrován s několika trénovanými klasifikátory obrázků tak, aby bylo možno extrahovat jak textové položky, tak obrázky. Přístup je demonstrován a hodnocen na úloze extrakce popisů bicyklů nabízených různými internetovými obchody. Popsáno je i několik algoritmů pro klasifikaci obrázků s použitím různých množin rysů pro klasifikaci.

Tyto trénované přístupy jsou posléze integrovány v rámci navržené metody *rozšířených extrakčních ontologií*, navazující na práci D.W. Embleyho [10], kterou rozšiřuje o současné využití všech tří výše zmíněných typů extrakční znalosti. Zamýšlenými přínosy extrakčních ontologií jsou rychlý vývoj funkčního prototypu, jeho plynulý přechod do finální IE aplikace a možnost využít různá množství tří typů extrakční znalosti podle jejich dostupnosti.

Protože extrakční ontologie je typicky odvozena z vhodné doménové ontologie a zůstává ve středu extrakčního procesu, minimalizuje tento přístup úsilí nutné pro zpětnou konverzi extrahovaných výsledků pro populaci zdrojové ontologie či datového schématu. Výsledky navrženého přístupu jsou prezentovány pro několik reálných domén.

Klíčová slova: extrakční ontologie, extrakce informací, skryté markovské modely, extrakce obrázků, klasifikace obrázků.

Abstract and keywords

Automatic information extraction (IE) from various types of text became very popular during the last decade. Owing to information overload, there are many practical applications that can utilize semantically labeled data extracted from textual sources like the Internet, emails, intranet documents and even conventional sources like newspaper and magazines.

Applications of IE exist in many areas of computer science: information retrieval systems, question answering, machine translation or website quality assessment. This work focuses on developing IE methods and tools that are particularly suited to extraction from semi-structured documents such as web pages and to situations where available training data is limited.

The main contribution of this thesis is the proposed approach of *extended extraction ontologies*. It attempts to combine extraction evidence from three distinct sources: (1) manually specified extraction knowledge, (2) existing training data and (3) formatting regularities that are often present in online documents.

The underlying hypothesis is that using extraction evidence of all three types by the extraction algorithm can help improve its extraction accuracy and robustness, and shorten the path to a functional prototype. The motivation for this work has been the lack of described methods and tools that would exploit these extraction evidence types at the same time.

This thesis first describes a statistically trained approach to IE based on Hidden Markov Models which integrates with a picture classification algorithm in order to extract product offers from the Internet, including textual items as well as images. This approach is evaluated using a bicycle sale domain. Several methods of image classification using various feature sets are described and evaluated as well.

These trained approaches are then integrated in the proposed novel approach of *extended extraction ontologies*, which builds on top of the work of D.W. Embley [10] by exploiting manual, trained and formatting types of extraction evidence at the same time. The intended benefit of using extraction ontologies is a quick development of a functional IE prototype, its smooth transition to deployed IE application and the possibility to leverage the use of each of the three extraction evidence types.

Also, since extraction ontologies are typically developed by adapting suitable domain ontologies and the ontology remains in center of the extraction process, the work related to the conversion of extracted results back to a domain ontology or schema is minimized. The described approach is evaluated using several distinct real-world datasets.

Keywords: extraction ontologies, information extraction, Hidden Markov Models, image extraction, image classification.

1 Cíle disertační práce

Předkládaná práce se zabývá metodami extrakce informací z dokumentů obsahujících text a obrázky. Vyvinuty a popsány jsou nové metody extrakce informací, které rozšiřují existující přístupy o nové prvky, nebo tyto přístupy kombinují. Navržené metody byly implementovány v různých programovacích jazycích a jejich výsledky byly vyhodnoceny na reálných datech a jsou prezentovány v práci.

1.1 Hlavní cíl disertační práce

Hlavním cílem práce je analyzovat možnosti využití různých typů znalostí pro extrakci informací z dokumentů. Práce rozšiřuje koncept tzv. *extrakčních ontologií*, původně navržených týmem D. W. Embleyho [10], o možnost zapojení tří hlavních typů extrakčních znalostí – od lidského experta, indukovaných z trénovacích dat, a nesupervizovaně indukovaných z pravidelné formátovací struktury zpracovávaných dokumentů. Motivací pro výběr tohoto tématu byl nedostatek přístupů a nástrojů, které by umožňovaly současné využití všech tří uvedených typů extrakčních znalostí. Hypotézou, kterou se v práci snažím potvrdit, je výhodnost tohoto přístupu oproti v praxi často používaným metodám, které se spoléhají čistě na trénovací data, čistě na ručně zadaná pravidla nebo na specifické formátování určité skupiny webových stránek.

V práci je navržen jazyk *EOL (Extraction Ontology Language)*, sloužící k zachycení *extrakčního schématu* a *extrakčních znalostí* výše uvedených typů. V rámci práce byl rovněž implementován nástroj *Ex* pro extrakci pomocí *rozšířených extrakčních ontologií*, aby bylo možno verifikovat navržené přístupy na reálných úlohách a změřit a zhodnotit jejich extrakční výsledky.

1.2 Dílčí cíle disertační práce

- Navržení, implementace a evaluace několika klasifikátorů obrázků za účelem obohacení extrakce textových položek o extrakci relevantních obrázků.
- Implementace extrakčního nástroje založeného na skrytých markovských modelech (HMM), návrh, implementace a experimenty s třemi různými variantami modelu.
- Rozšíření HMM extrakce o extrakci obrázků pomocí klasifikátorů obrázků.
- Implementace v praxi využitelného open-source extrakčního nástroje *Ex* podporujícího v práce navržené rozšířené extrakční ontologie.
- Podat ucelený přehled o aktuálních metodách vyžívaných pro extrakci informací.

2 Metody řešení a dosažení stanovených cílů

Protože většina cílů práce má praktický charakter a jejich řešení vyžaduje provádění experimentů, zvolil jsem pro řešení většiny z nich následující postup:

1. detailní vymezení problému,
2. analýza existujících publikovaných metod řešení,
3. návrh metody řešení (návrh či výběr algoritmů, klasifikátorů apod.),
4. implementace a případné iterace zpět k bodům 2 nebo 3,
5. evaluace na reálných datech a případné iterace zpět k bodům 2, 3 nebo 4.

2.1 Rozšířené extrakční ontologie

Návrhu a vývoji rozšířených extrakčních ontologií, které jsou hlavní náplní práce, předcházela analýza požadavků na cílový extrakční systém. Mezi stěžejními požadavky byla možnost rychlého prototypování extrakčních úloh, snadné změny extrakčního schématu a schopnost vyvažovat nákladnost a přesnost extrakce. Pro splnění těchto požadavků se ukázalo jako velmi užitečné využít všech tří typů extrakčních znalostí zmíněných v předchozí kapitole.

Jako výchozí bod pro výzkum rozšířených extrakčních ontologií byla zvolena práce D.W. Embleyho [10], která definuje základní koncept extrakčních ontologií s aplikací pro extrakci položek z produktových katalogů na Internetu. *Rozšířené extrakční ontologie*, navržené v této disertační práci, tento koncept rozšiřují zejména o:

- využití tří typů extrakční znalosti,
- navržený jazyk *EOL* pro reprezentaci extrakčních ontologií,
- spojení extrakčních znalostí s pseudo-pravděpodobnostními parametry a metodu jejich kombinace,
- algoritmy pro identifikaci kandidátů na hodnoty atributů, pro generování kandidátů na instance, pro indukci formátovacích vzorů a pro nalezení nejlepší sekvence extrahovaných položek v dokumentu.

Navržený přístup byl implementován v podobě open-source extrakčního nástroje *Ex*[1]. Za účelem evaluace na reálných datech byl tento systém doplněn o řadu podpůrných procesů, jako je evaluační komponenta pro měření přesnosti extrakce různými metodami, podpora pro křížovou validaci, vstup a výstup v různých formátech, integrace s dalšími extrakčními nástroji apod. Systém je implementován v jazyce Java v rozsahu cca 50,000 řádků zdrojového kódu.

Pro ověření využitelnosti rozšířených extrakčních ontologií a nástroje *Ex* byly provedeny experimenty na čtyřech reálných doménách.

1 Binární distribuce systému Ex, zdrojové kódy včetně sestavovacího systému, příklady extrakčních ontologií pro 3 domény a tutoriál pro vývoj extrakčních ontologií jsou k dispozici na http://eso.vse.cz/~labsky/ex

2.2 Skryté markovské modely

Experimenty se *skrytými markovskými modely* (Hidden Markov Models, HMM) popsané v práci vycházejí z prací D. Freitaga a A. Mc. Calluma [11]. Cílem bylo ověřit aplikovatelnost HMM na doménu produktových katalogů určitého typu zboží na Internetu. Motivací byl fakt, že HMM nebyly dosud pro tuto doménu využívány a pokud byly aplikovány na webové stránky, během předzpracování byla z dokumentů typicky odstraněna informace o formátovací struktuře. Přístup popsaný v disertační práci:

- modeluje pomocí HMM text včetně předzpracované formátovací struktury,
- využívá pro extrakci mnoha typů položek jediný HMM model, čímž se odlišuje od [11] a je podobný [3],
- experimentuje se třemi různými variantami HMM modelu – (1) naivní, kde obsah každé extrahované položky je modelován unigramovou lexikální distribucí jediného stavu, (2) modifikovanou, kde je unigramová lexikální distribuce nahrazena trigramovou a (3) topologií skládající se z indukovaných submodelů pro každou extrahovanou položku, které její obsah modelují pomocí několika HMM stavů,
- využívá trigramový HMM model.

Tyto experimenty s HMM časově předcházely vývoji extrakčních ontologií a v po vzniku systému *Ex* byl vyvinutý HMM nástroj integrován jako jeden z trénovaných algoritmů dostupný v rámci extrakčních ontologií.

2.3 Extrakce obrázků

Extrakce informací z produktových katalogů, na niž byly aplikovány HMM modely z předešlé sekce, vyžadovala rovněž extrakci obrázků nabízeného zboží. Za tímto účelem byly v práci vyvinuty binární klasifikátory obrázků, trénované tak, aby rozlišily relevantní obrázky zboží od všech ostatních obrázků. Úloha byla zjednodušena tím, že zpracovávány byly vždy pouze dokumenty z dané domény.

Po analýze publikovaných přístupů k podobným klasifikačním úlohám byla definována množina rysů, které mohly být pro klasifikaci užitečné. Patřily mezi ně rozměry obrázku, předzpracovaný barevný histogram, počet výskytů téhož obrázku ve stránce a míra podobnosti obsahu obrázku k trénovací kolekci, vyvinutá kolegy z VŠB TU Ostrava [28]. Protože bylo zajímavé zjistit příspěvky jednotlivých rysů ke klasifikaci, byly zkonstruovány jak klasifikátory využívající celou množinu rysů, tak její části.

Výsledný klasifikátor obrázků byl dále integrován v rámci výše popsaného HMM modelu tak, že lexikální distribuce HMM stavů generovaly vedle slov a formátovacích symbolů také třídy obrázků, které měl klasifikátor obrázků na výstupu. Experimenty byly provedeny s ternární variantou klasifikátoru, který klasifikoval obrázky jako pozitivní (relevantní obrázky produktů), negativní a nejisté. Pro obrázky klasifikované jako nejisté zde měl hrát větší roli kontext jejich výskytu.

3 Stav řešení problematiky v ČR a ve světě

V této kapitole nejprve uvedeme členění metod IE z tří různých hledisek, poté popíšeme aktuální trendy v IE a nakonec zmíníme některé aplikace a výzkumné aktivity týkající se IE v ČR. Přehledová studie [14] obsahuje popis celé řady metod a především nástrojů pro IE, včetně jejich vzájemných srovnání. Protože je IE stále ještě relativně nová oblast, není k dispozici mnoho knižních publikací na toto téma. Výjimkou je monografie [26], která obsahuje úvod do problematiky IE a její evaluace, popisuje některé stěžejní IE metody a algoritmy a zmiňuje několik praktických aplikací.

3.1 Členění IE metod

IE metody lze členit dle několika kritérií. Nejprve uvedeme členění IE podle jejích podúloh:

- *Extrakce pojmenovaných entit (named entity extraction)* je základní disciplínou, jejímž cílem je nalézt v textu relevantní textové hodnoty, např. jména osob, časové údaje nebo názvy filmů.
- *Analýza koreferencí (coreference resolution)* se snaží nalézt v textu odkazy, např. mezi zájmenem a jménem nebo několika pojmenovanými entitami odkazujícími na stejný reálný objekt.
- *Populace šablon (template filling)* je úloha, která má za cíl naplnit extrahovanou informací šablony, popisující nějaký typ události či objektu. Např. může jít o extrakci času, místa a druhu teroristických činů z novinových článků.
- *Extrakce relací (relation extraction)* se snaží nalézt mezi extrahovanými položkami (mohou jimi být jak pojmenované entity tak celé instance šablon) jeden z předem daných vztahů, např. „zaměstnanec pracuje pro firmu.“

Disertační práce se věnuje prvním třem disciplínám. Další vhodné členění je možné podle typu zpracovávaných textů:

- *IE z přirozeného textu* často využívá celé řady podpůrných lingvistických nástrojů.
- *IE z webových stránek* se typicky nespoléhá na přirozený gramatický text, ale může využít formátovací struktury přítomné na stránkách nebo skupinách stránek. Disertační práce se věnuje především této kategorii.
- *IE z databázových záznamů* mívá za úkol *integraci informací*, např. může jít o extrakci jednotlivých položek volně zadaných adres za účelem odstranění duplicit.

Kategorizace dle typu extrakčního algoritmu rozdělí IE metody na:

- *Manuální techniky* jsou často založeny na kaskádách ručně zadaných pravidel (např. [13]), které mohou využívat např. regulární výrazy nebo odkazovat na předpokládanou formátovací strukturu (např. [16] a částečně [1]).
- *Trénovatelné techniky* jsou schopny učit se (semi-)automaticky z dat. Většina prakticky používaných metod je supervizovaných, tj. vyžadují trénovací data v podobě textů, kde jsou ručně anotovány položky, které je třeba extrahovat. Učící se

algoritmus na základě těchto příkladů generalizuje a je proto schopen pokrýt i předtím neviděné příklady. Učící se IE algoritmy můžeme dále kategorizovat jako standardní algoritmy strojového učení:

- *symbolické* (např. pokrývání množin, indukce rozhodovacích stromů [2]),
- *pravděpodobnostní* (např. HMM [11] a CRF [24]),
- *ostatní subsymbolické* (např. neuronové sítě [7] nebo SVM [32]).

Mezi aktuální trendy v IE patří techniky minimalizující množství trénovacích dat pro učící se IE algoritmy. Prvním takovým přístupem je tzv. *bootstrapping*, který umožňuje využít iniciální funkční systém (získaný např. pomocí minima trénovacích dat nebo na základě manuálních pravidel) k automatické anotaci velkého množství dat. Jedná se o nesupervizovanou techniku, která na základě doménově závislých heuristik odvodí z automaticky anotovaných dat nové extrakční znalosti, které přidá k aktuálnímu extrakčnímu modelu. Ten je v další iteraci opět použit na další data a v ideálním případě tak dochází ke zlepšování výsledků modelu. Možnosti využití této metody pro IE jsou rozebrány v [35], [5] a [9].

Aktivní učení (active learning) je metoda, při které učící se algoritmus ve fázi tréninku interaktivně spolupracuje s lidským anotátorem, kterému předkládá k ruční anotaci vždy ten dokument, u něhož si je nejméně jistý svou automatickou anotací. Předkládaný dokument je navíc vždy předvyplněn navrženou automatickou anotací a uživatel jen opravuje chyby. Učení zde probíhá inkrementálně s každým novým ručně anotovaným dokumentem. Příkladem tohoto přístupu je systém *Melita* [6].

3.2 Výzkum IE v ČR

Na akademických pracovištích v ČR je obecně velký zájem o vývoj metod automatického zpracování přirozených textů (psaných i mluvených) a extrakce informací je jen jedním z těchto oborů. Aplikace rozpoznávání pojmenovaných entit pro podporu systému strojového překladu [36] je popsána v [18]. Metoda pro nesupervizované získávání extrakčních vzorů z webu je navržena v [17]. Lingvistická analýza a následná aplikace extrakčních pravidel za účelem extrakce popisů událostí z přirozených textů je řešena v [8]. Semi-automatický nástroj pro anotaci lékařských zpráv je popsán v [37]. Extrakce informací z webu s využitím ustálených návrhových vzorů je navržena v [19]. Přístup k IE z formátovaných dokumentů na základě propojení logické hierarchie prezentovaných dat a hierarchie formátovacích prvků je popsán v [4].

4 Struktura a obsah práce

První dvě kapitoly práce obsahují úvod do problematiky a popisují současný stav extrakce informací. Následující čtyři kapitoly popisují výzkum a experimenty provedené v oblasti IE pomocí v práci vyvinutých metod a nástrojů.

Kapitola 1 obsahuje stručný úvod do problematiky, vytyčuje cíle a vysvětluje motivaci výzkumu. Kapitola 2 popisuje současný stav extrakce informací, uvádí členění IE disciplín, diskutuje druhy IE podle typu zpracovávaných dat, člení a stručně popisuje často používané algoritmy a metody.

Kapitola 3 popisuje vývoj binárních klasifikátorů obrázků a výsledky klasifikačních experimentů na reálných obrázcích z domény cyklistických produktů. Upravený klasifikátor z této kapitoly je použit v následující kapitole 4 o IE pomocí HMM. Tato kapitola popisuje vývoj HMM extrakčního nástroje s využitím open-source POS taggeru [30], který je zde rozšířen pro potřeby IE. Jsou zde popsány experimenty s třemi variantami HMM modelu a způsob integrace klasifikátoru obrázků s HMM extrakčním modelem. Výsledky jsou opět prezentovány pro produktové katalogy z cyklistické domény.

Kapitola 5 definuje rozšířené extrakční ontologie. Začíná popisem navrženého a implementovaného extrakčního jazyka EOL, vysvětluje jeho použití pro definici extrakčního schématu a pro zadávání extrakčních znalostí tří typů – ručně zadaných, indukovaných z trénovacích dat, a indukovaných z pravidelné formátovací struktury. Dále je popsán model, podle kterého se extrakční znalosti (evidence) kombinují za zjednodušujících předpokladů nezávislosti. Nakonec je popsán extrakční proces, sestávající z několika fází, pro které jsou navrženy a implementovány extrakční algoritmy. Open-source extrakční nástroj *Ex*, vyvinutý v rámci této kapitoly, je prakticky využíván pro extrakci kontaktních a dalších údajů z webových stránek v medicínské doméně. Kapitola 6 popisuje experimentální výsledky extrakčních ontologií pro čtyři různé domény – e-mailová oznámení o seminářích, kontaktní informace z webových stránek, produktové katalogy z cyklistické domény (s využitím HMM extrakce z kapitoly 4) a katalogy počítačových monitorů a televizí. Z těchto případových studií je na závěr kapitoly odvozena doporučená metodologie pro využití extrakčních ontologií v praktických IE projektech.

Kapitola 7 uzavírá práci rekapitulací vyvinutých metod a cílů práce, a uvádí možné směry budoucího výzkumu.

Obsah disertační práce je uveden jako příloha A tohoto autoreferátu.

5 Použitá terminologie

Atribut (z extrakční ontologie). Základní extrahovatelná položka *extrakční ontologie*. Může odpovídat jedné *pojmenované entitě* v širším slova smyslu (např. jméno osoby, telefonní číslo), ale také může modelovat delší souvislý text (např. odstavec určitého typu). Skupiny atributů mohou tvořit *třídy* extrakční ontologie.

Conditional Random Field (CRF). Statistická metoda učící se z trénovacích dat, využívaná v extrakci informací pro nalezení nejpravděpodobnější sekvence sémantických značek pro vstupní sekvenci slov. Při tréninku CRF iterativně odhadují podmíněné pravděpodobnostní distribuce, modelující pravděpodobnost předpovídaných sémantických značek, dáno pozorované hodnoty rysů. Tím se odlišují od generativních metod, jakými jsou např. HMM, a umožňují zapojení velkého počtu heterogenních rysů, které mohou být vzájemně závislé. Více informací viz. [24] a [20].

Ex (software). Open-source nástroj pro extrakci informací pomocí *rozšířených extrakčních ontologií*, implementovaný v rámci disertační práce, popsaný v [21] a dostupný na http://eso.vse.cz/~labsky/ex [21].

Extraction Ontology Language (EOL). Jazyk pro reprezentaci rozšířených extrakčních ontologií, navržený a implementovaný v disertační práci. Zachycuje jak extrakční schéma, tak znalosti potřebné pro identifikaci extrahovaných položek v textu.

Extrakce informací (Information Extraction, IE). Automatizovaný proces, jehož cílem je identifikace informací předem definovaného významu v analyzovaných dokumentech (např. jméno přednášejícího, místo a čas konání semináře z relevantních emailových oznámení).

Extrakční ontologie. *Extrakční schéma* doplněné o *extrakční znalosti*. Termín zavedl D.W. Embley [10]. V širším smyslu označuje nejen způsob reprezentace znalostí potřebných k extrakci, ale také extrakční metodu (použité algoritmy).

Extrakční schéma. Množina extrahovatelných *atributů* a *tříd*. Lze vyjádřit např. E-R diagramem nebo UML modelem tříd.

Extrakční znalosti. Znalosti umožňující identifikaci informací relevantních pro extrakci ve vstupních dokumentech. Mohou být vyjádřeny v různých formách, např. pravidla využívající regulární výrazy nebo natrénovaný pravděpodobnostní model. Mohou být rovněž různého původu – ručně zadané, indukované z ručně anotovaných trénovacích dokumentů, nebo nesupervizovaně indukované ze zpracovávaných dokumentů.

Hodnota atributu (z extrakční ontologie). Konkrétní naplnění extrahovatelného *atributu*. V *rozšířených extrakčních ontologiích* jde konkrétně o finální extrahovanou hodnotu po případné aplikaci post-extrakčních pravidel, které mohou extrahované textové položky upravovat (např. převést datum do jednotného formátu).

Ontologie. V oblasti informačních systémů je ontologie chápána jako sdílená a formalizovaná konceptualizace určitého oboru lidské činnosti [12]. Obsahuje formální popis relevantních konceptů vybrané domény, jejich hierarchii a další druhy relací mezi nimi. Může umožňovat strojové odvozování nad těmito znalostmi.

Instance (třídy z extrakční ontologie). Konkrétní výskyt instance extrahovatelné *třídy*. Např. při extrakci oznámení o seminářích může jít o k sobě patřící čtveřici hodnot atributů *přednášející*, *místo*, *čas začátku* a *konce semináře*.

Part-of-speech tagger (POS tagger). Automatický nástroj sloužící k přiřazení různých lingvistických kategorií ke slovům v analyzovaném textu. Základní kategorií je slovní druh (např. podstatné jméno vlastní). V češtině může jít o vysoký počet kategorií z nichž většina je relevantní pouze pro určité slovní druhy (např. pád, osoba, čas). Implementace *POS taggerů* pomocí *HMM* a dalších metod je popsána např. v [30].

Pojmenovaná entita (named entity, NE). Základní extrahovatelná položka v *extrakci informací*. V užším smyslu se *pojmenovanou entitou* rozumí např. jména osob, časové údaje nebo geografické názvy. V širším smyslu jde o krátké textové položky libovolného významu. V extrakčních ontologiích jsou pojmenované entity modelovány extrahovatelnými *atributy*.

Rozšířená extrakční ontologie. Rozšíření pro *extrakční ontologie* vyvinuté v rámci disertační práce. Zahrnuje zejména: využití tří typů *extrakčních znalostí* pro extrakci (ručně zadané, indukované z trénovacích dat, nesupervizovaně indukované z pravidelné formátovací struktury), jazyk *EOL* pro reprezentaci *extrakčních ontologií*, metodu kombinace *extrakčních znalostí* a extrakční algoritmy.

Skrytý markovský model (Hidden Markov Model, HMM). Statistická metoda učící se z trénovacích dat, využívaná v extrakci informací pro nalezení nejpravděpodobnější sekvence sémantických značek pro vstupní sekvenci slov. *HMM* modeluje dokument generativním způsobem. Je reprezentován pravděpodobnostním konečným automatem, jehož stavy (nebo přechody) emitují symboly (typicky slova dokumentu). Pro účely IE vybrané stavy modelují položky určené k extrakci, zatímco jiné stavy modelují irelevantní zbytek dokumentu. Při tréninku z ručně anotovaných dat jsou odhadovány lexikální pravděpodobnostní distribuce, s nimiž jsou z určitého stavu (nebo přechodu) emitovány konkrétní symboly (slova), a přechodové distribuce mezi jednotlivými stavy. Při vlastní extrakci je použit Viterbi algoritmus [29] pro nalezení nejpravděpodobnější cesty automatem, která měla vygenerovat text analyzovaného dokumentu, z jejíchž stavů lze odvodit výstupní sémantické značky.

Třída (z extrakční ontologie). Skupina extrahovatelných atributů tvořících logický celek. Např. třída „oznámení o semináři“ seskupuje atributy jako jsou jméno přednášejícího, místo a čas konání.

6 Zhodnocení práce

Následující tři sekce shrnují výsledky dosažené v oblasti výzkumu rozšířených extrakčních ontologií, výsledky experimentů s HMM pro IE z webových produktových katalogů a výsledky extrakce obrázků.

6.1 Rozšířené extrakční ontologie

V práci byla vyvinuta metoda IE, umožňující snadné kombinování ručně zadaných extrakčních znalostí, trénovacích dat a znalostí indukovaných z pravidelné formátovací struktury. Dále byl navržen jazyk EOL pro reprezentaci rozšířených extrakčních ontologií a vyvinut open-source nástroj pro IE, který navržené extrakční ontologie implementuje. Výsledky tohoto výzkumu byly publikovány mj. v [21].

Využitelnost přístupu byla verifikována na čtyřech reálných doménách. Výsledky extrakce oznámení o seminářích jsou uvedeny v Tabulce 1, která porovnává extrakční ontologii využívající ručně zadané extrakční znalosti s ontologií, která tyto znalosti kombinuje s trénovaným Conditional Random Field (CRF) modelem [20]. Výsledky ukazují signifikantní zlepšení F-míry pro kombinovanou extrakční ontologii; konfidenční intervaly pro hladinu pravděpodobnosti α=0.5 pro F-míry byly 79.56±1.57 vs. 87.6±1.42.

	manuální EO, testovací data			kombinovaná EO, 10-CV			shrnutí	
atribut	přesnost	úplnost	F-míra	přesnost	úplnost	F-míra	rozdíl F	počet
Speaker	69.9	66.5	68.1	75.4	75.0	75.2	+7.1	689
– loose	76.2	72.7	74.4	81.8	80.6	81.2	+6.8	
Location	59.7	75.9	66.9	93.3	78.0	85.0	+18.1	575
– loose	77.5	86.0	81.5	97.6	80.7	88.3	+6.8	
Start time	96.0	88.7	92.2	98.1	93.3	95.6	+3.4	881
– loose	96.4	88.9	92.5	98.1	93.3	95.6	+3.1	
End time	97.8	90.3	93.9	97.0	94.4	95.7	+1.8	380
– loose	97.9	90.5	94.1	97.2	94.7	96.0	+1.9	
Celkem	79.1	80.0	79.6	90.4	85.0	87.6	+8.0	2525
– loose	85.9	84.1	85.0	93.2	87.2	90.1	+5.0	

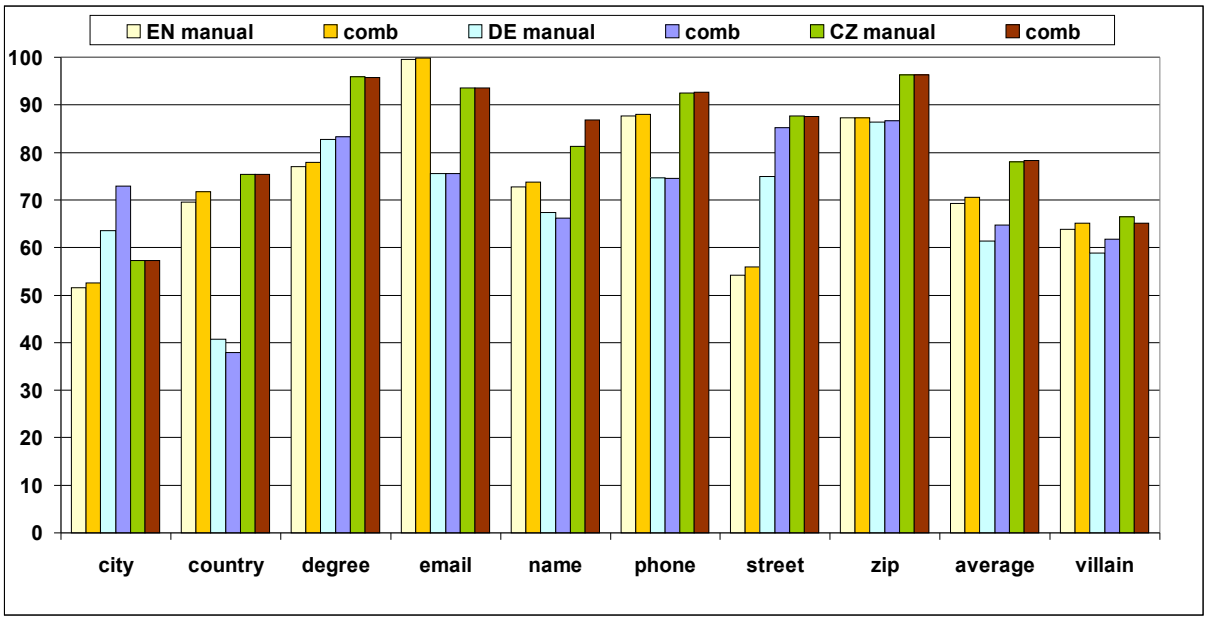
Tabulka 1. Výsledky extrakce z emailových oznámení o seminářích pomocí ručně zadaných znalostí a pomocí kombinace těchto znalostí s trénovaným CRF modelem.

Tabulka 2 uvádí pro úlohu extrakce seminářů srovnání s publikovanými výsledky dalších nástrojů. Nejlépe se umístila bayesovská síť BIEN využívající rozsáhlou ručně definovanou množinu příznaků. Výkon extrakčních ontologií je cca o 2 procentní body nižší, vzhledem k relativně malé velikosti analyzovaných dat jde však o rozdíl těsně za hranicí statistické signifikance. Je nutné dodat, že LP2, SRV, Rapier a Whisk nevyužívaly ručně zadané znalosti ani příznaky; srovnání s nimi je tedy pouze orientační.

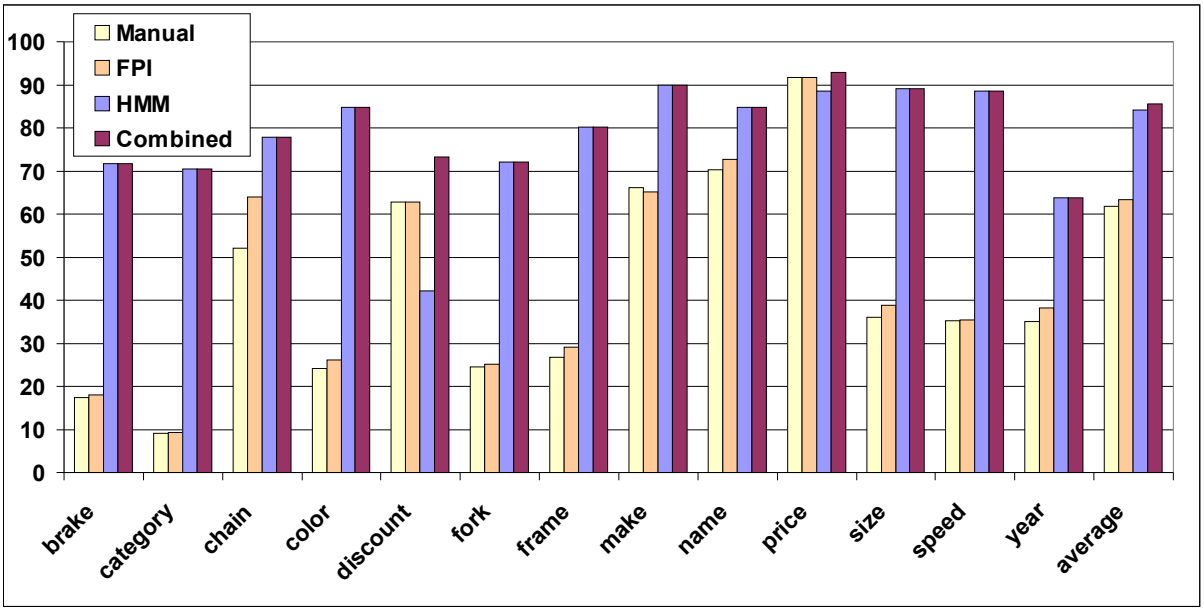
	BIEN	LP2	EO	SRV	Rapier	Whisk
Speaker	76.9	**77.6**	75.2	66.2	53.0	18.3
Location	**87.1**	75.1	85.0	79.7	73.3	66.4
Start time	96.0	**99.0**	95.6	94.3	95.9	92.6
End time	**98.8**	95.5	95.7	99.3	96.7	86.1
Overall	-	89.9	87.6	-	82.6	-

Tabulka 2. Srovnání výsledků extrakce z emailových oznámení o seminářích pro různé extrakční nástroje. Extrakční ontologie jsou označeny *EO,* nejlepší F míry tučně.

Systém byl dále úspěšně aplikován v projektu MedIEQ [25], zabývajícím se automatickou podporou doposud ručního anotování kvality medicínských webových stránek, kde byl využit pro extrakci kontaktních informací a několika volno-textových kritérií. Výsledky extrakce pro 11 atributů (pojmenovaných entit), dosažené pro 3 jazyky vždy pomocí dvou extrakčních ontologií, jsou prezentovány na Obr. 1. Atributy *department*, *job* a *organization* vykazovaly v anotovaných kolekcích dokumentů velmi nízkou míru anotátorské shody nebo nebyly anotovány, pro úplnost jsou však tyto atributy uvedeny také. Graf srovnává pro každý atribut a jazyk F-míru dosaženou pomocí extrakční ontologie obsahující jen ručně zadané znalosti, a pomocí ontologie kombinující tyto znalosti s trénovaným CRF modelem a s indukcí formátovacích vzorů. Výsledky kombinované varianty zde vykazují jen malé zlepšení, což může být zapříčiněno mj. nižší anotátorskou shodou v trénovacích datech.



Obr. 1. Výsledky (F-míry) extrakce kontaktních informací z medicínských webových stránek pro angličtinu, němčinu a češtinu pomocí manuálních a kombinovaných extrakčních ontologií. Villain score hodnotí přesnost seskupování extrahovaných položek do instancí.



Obr. 2. Výsledky (F-míry) extrakce popisů bicyklů z webových stránek prodejců v Anglii pomocí čtyř extrakčních ontologií: (1) obsahující pouze ručně zadané extrakční znalosti, (2) využívající navíc detekci pravidelného formátování, (3) spoléhající čistě na trénovaný HMM model a datotypová omezení atributů, (4) kombinující výše uvedené znalosti.

Další testovanou doménou byla extrakce popisů bicyklů v katalozích internetových obchodů v Anglii. Obr. 2 ukazuje výsledky (F-míry) čtyř extrakčních ontologií pro tuto doménu. Výsledky čistě manuální ontologie ontologie nejsou pro většinu atributů uspokojivé. K mírnému zlepšení dochází pro některé atributy po zapnutí nesupervizované indukce formátovacích vzorů. Ontologie založená na trénovaném HMM modelu je v průměru výrazně lepší, nicméně pro některé atributy dochází k poklesu F-míry (*discount*, *price*). Nejlepšího výsledku zde dosahuje extrakční ontologie kombinující všechny tři zdroje extrakčních znalostí. Detailní podmínky experimentů jsou popsány v disertační práci. V práci je dále popsána související aplikace extrakčních ontologií pro extrakci z produktových katalogů počítačových monitorů a televizí.

Popsané experimenty na reálných datech indikují využitelnost rozšířených extrakčních ontologií diky následujícím hlavním faktorům:

- snadná kombinace tří zdrojů extrakčních znalostí, vedoucí ke zlepšení přesnosti výsledků a k rychlejšímu a levnějšímu vývoji IE prototypu,
- snadné změny extrakčního schématu,
- nižší nároky na zpětnou integraci extrahovaných dat do ontologie.

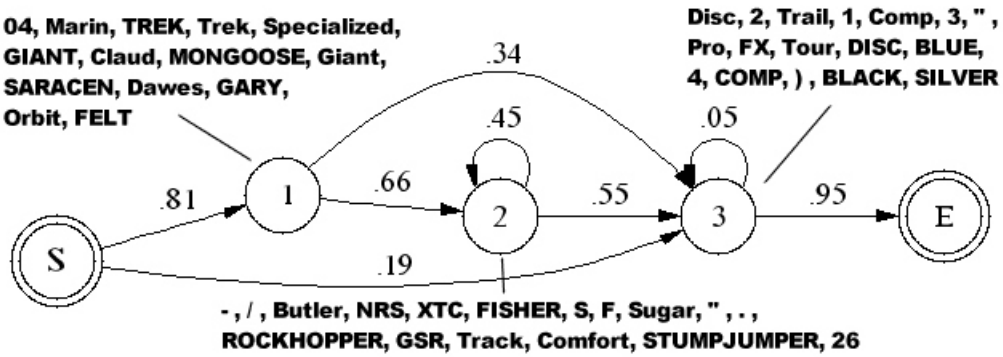
6.2 Skryté markovské modely

Experimenty s HMM modely pro extrakci z webových stránek vedly k několika závěrům:

- HMM jsou dobře využitelné pro extrakci z webových produktových katalogů v případě dostatku kvalitních trénovacích dat.
- Přínos složitějších topologií HMM modelu, oproti základnímu naivnímu modelu,

není výrazný pro trénovací data v rozsahu stovek stránek a jednotek tisíců pojmenovaných entit. Nevýrazný byl také přínos zavedení n-gramů (kde n > 1) pro modelování lexikálních emisních pravděpodobnostní sekvencí slov generovaných v rámci jednoho HMM stavu.

- Nesupervizovaná indukce topologie HMM modelů pro obsah jednotlivých pojmenovaných entit EM (Baum-Welch) algoritmem [15] vede k modelům, kde vytvořené stavy velmi dobře odpovídají sémantickému členění typických obsahů dané pojmenované entity. Obr. 3 ukazuje indukovaný model pro název modelu bicyklu.
- Byla navržena a úspěšně implementována metoda pro integraci HMM modelu s klasifikátorem obrázků. HMM lze snadno upravit pro extrakci netextových položek tak, že do lexikálních symbolů zahrneme i třídy produkované zúčastněnými klasifikátory a při tréninku i testování před aplikací HMM nahradíme originální obsah v dokumentech identifikátory predikovaných tříd.
- Byl implementován HMM extrakční nástroj (rozšířením POS taggeru [30] pro potřeby IE) který byl použit pro vývoj doménově specifického strukturovaného vyhledávače popsaného v [22]. HMM nástroj byl následně pro experimenty s extrakčním ontologiemi integrován do systému *Ex*.



Obr. 3. HMM model indukovaný nesupervizovaným EM (Baum-Welch) algoritmem z pozitivních trénovacích příkladů názvů modelů bicyklů. Indukované stavy přibližně korespondují s (1) jménem výrobce, (2) jménem modelu, (3) doplňující informací. Zobrazeny jsou jen nejpravděpodobnější slova emitovaná každým stavem a přechody pravděpodobnější než 0.05 (včetně přechodových pravděpodobností). Stavy *S* a *E* jsou počáteční resp. koncové.

6.3 Extrakce obrázků

Kapitola 4 disertační práce popisuje několik vyvinutých variant binárních klasifikátorů obrázků využívajících pro klasifikaci různé rysy. Úkolem je odlišit relevantní obrázek od irelevantního, pro naši doménu obrázek bicyklu od libovolné jiné grafiky. Tabulka 2 shrnuje chybovost jednotlivých klasifikátorů. Přesnost klasifikace je zde relativně vysoká, protože všechny klasifikované obrázky pocházejí z webových stránek patřících do zkoumané domény. Negativní příklady jsou tedy tvořeny především doprovodnou grafikou na webových stránkách, a pouze z malé části obrázky produktů irelevantního typu.

Převzatá míra podobnosti obsahu obrázků, vyvinutá autory [28] a zde aplikovaná jako podobnost obrázku ke skupině pozitivních trénovacích příkladů (*Podobnost*), nevedla sama a sobě k uspokojivé přesnosti. Překvapivě dobré výsledky však byly dosaženy pomocí pouhých rozměrů obrázků. Klasifikátor označený *Rozměry 1* klasifikoval pouze na základě hodnoty 2-dimenzionální hustoty pravděpodobnosti, která byla odhadnuta z horizontálních a vertikálních rozměrů (v pixelech) trénovacích pozitivních příkladů. Klasifikátor *Rozměry 2* dosáhl ještě lepšího výsledku dodatečným zahrnutím absolutních hodnot rozměrů jako rysů pro klasifikaci. Rysy rovněž úspěšného klasifikátoru nazvaného *Histogram* byly tvořeny předzpracovaným barevným histogramem obrázku. Klasifikátor označený *Kombinace* byl natrénován jako rozhodovací seznam PART [34] s využitím všech zmíněných rysů a dosáhl nejvyšší přesnosti.

	Podobnost	Rozměry 1	Rozměry 2	Histogram	Kombinace
Chybovost (%)	26.4	6.2	3.2	5.2	2.6

Tabulka 2. Chybovost klasifikátorů obrázků bicyklů využívajících různé rysy.

Úloha extrakce obrázků prodávaného zboží se ukázala být výrazně složitější než samotná klasifikace obrázku. Ne všechny obrázky bicyklů ve zkoumané kolekci stránek byly totiž označeny pro extrakci: řada obrázků kol byla na stránce pouze v roli ilustrace a nebyla součástí konkrétní nabídky, která by byla předmětem extrakce. Pro extrakci byl použit naivní HMM model, do kterého byl integrován výše uvedený kombinovaný klasifikátor, upravený tak, aby navíc klasifikoval také do třídy „nevím“ a dal tak HMM modelu možnost dát pro nejisté případy větší váhu kontextu, ve kterém se obrázek vyskytoval. Tabulka 3 uvádí výsledky extrakce pro relevantní obrázky bicyklů.

	Přesnost	Úplnost	F-míra
Relevantní obrázek kola	83.8	82.5	83.2

Tabulka 3. Extrakce relevantních obrázků bicyklů z webových stránek pomocí HMM integrovaného s kombinovaným ternárním klasifikátorem obrázků.

Popsané experimenty ukazují, že vlastní klasifikace obrázků v rámci omezené domény může dosahovat vysoké přesnosti i s využitím velmi omezené množiny rysů. Klasifikátory je možné dále integrovat s extrakčními algoritmy, které mohou z klasifikovaných obrázků dále vybírat výskyty relevantní pro extrakci, a tyto extrahovat spolu se souvisejícími textovými informacemi.

7 Výběr z použité literatury

[1] Baumgartner, R., Wien, D.T., Flesca, S., Gottlob, G.: Supervised Wrapper Generation with Lixto. In: Proc. *VLDB Demo* 2001.

[2] Berka, P.: Dobývání znalostí z databází. *Academia*, 2003. ISBN 80-200-1062-9.

[3] Borkar, V., Deshmukh, K., Sarawagi, S.: Automatic segmentation of text into structured records. In: Proc. *SIGMOD* 2001.

[4] Burget, R.: Visual HTML Document Modeling for Information Extraction. In Proc. *RAWS* 2005.

[5] Ciravegna, F., Dingli, A., Guthrie, D., Wilks, Y.: Integrating Information to Bootstrap Information Extraction from Web Sites. In: Proc. *IJCAI Workshop on Information Integration on the Web*, 2003.

[6] Ciravegna, F., Dingli, A., Iria, C., Wilks, Y.: Multi-strategy Definition of Annotation Services in Melita. In: Proc. *International Semantic Web Conference (ISWC)*, 2003.

[7] Collins, M., Roark, B.: Incremental Parsing with the Perceptron Algorithm. In: Proc. *ACL* 2004.

[8] Dědek, J., Vojtáš, P.: Linguistic Extraction for Semantic Annotation. In: *Intelligent Distributed Computing, Systems and Applications, Studies in Computational Intelligence*, vol. 162, 2008, p. 85-94.

[9] Dingli, A., Ciravegna, F., Guthrie, D., Wilks, Y.: Mining Web Sites Using Unsupervised Adaptive Information Extraction. In: Proc. *EACL* 2003.

[10] Embley, D.W., Tao, C., Liddle, S.W.: Automatically extracting ontologically specified data from HTML tables with unknown structure. In: Proc. *ER 2002*.

[11] Freitag, D., McCallum, A.: Information extraction with HMMs and shrinkage. In: *AAAI Workshop on Machine Learning for IE 1999*.

[12] Gruber, T.: A translation approach to portable ontology specifications. In: *Knowledge Acquisition*, Vol. 5, 1993.

[13] Hobbs, J.R., Appelt, D.E., Bear, J., Israel, D., Kameyama, M., Stickel, M., Tyson, M.: FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In: *Finite State Devices for Natural Language Processing*, MIT Press, Cambridge, 1997.

[14] Chang, C.H., Kayed, M., Girgis, M.R., Shaalan, K.: A Survey of Web Information Extraction Systems. In: *IEEE Transactions on Knwoledge and Data Engineering*, Vol. 18/10, 2006. ISSN: 1041-4347.

[15] Jelinek, F.: Statistical Methods for Speech Recognition. *The MIT Press*, Cambridge, Massachusetts, 1997.

[16] Kapow Technologies: Kapow RoboSuite Whitepaper. http://www.kapowtech.com.

[17] Kavalec, M., Svátek, V.: Information Extraction and Ontology Learning Guided by Web Directory. In: *ECAI Workshop on NLP and ML for ontology engineering*. Lyon 2002.

[18] Kravalová, J.: Named Entities Recognition in Czech. *KEG Seminar*, VŠE Praha, 2009. http://keg.vse.cz/_slides/kravalova.pdf.

[19] Kudělka, M., Snášel, V., Lehečka, O., El-Qawasmeh, E.: *Web content mining using web design patterns*. In: Proc. IRI 2008, p. 232-237.

[20] Kudo, T.: CRF++: Yet Another CRF toolkit. http://crfpp.sourceforge.net.

[21] Labský, M., Svátek, V.: Combining Multiple Sources of Evidence in Web Information Extraction. In: Foundations of Intelligent Systems. Toronto, Springer-Verlag, 2008, p. 471-476, ISBN 978-3-540-68122-9, ISSN 0302-9743.

[22] Labský, M., Svátek, V., Šváb, O., Praks, P., Krátký, M., Snášel, V.: Information Extraction from HTML Product Catalogues: from Sorce Code and Images to RDF. In: Web Intelligence 2005, Compiégne. Los Alamitos: IEEE, 2005, p. 401-404, ISBN 0-7695-2415-X.

[23] Labský, M., Vacura, M., Praks, P.: Web Image Classification for Information Extraction. In: RAWS 2005, VŠB TU, Ostrava, p. 55-62. ISBN 80-248-0864-1.

[24] Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. *18th International Conf. on Machine Learning*, 2001, p. 282-289.

[25] Mayer, M.A., Karkaletsis, V., Stamatakis, K., Leis, A., Gonzales, D.V., Thomeczek, C., Labský, M., López-Ostenero, F., Honkela, T.: MedIEQ – Quality Labelling of Medical Web Content Using Multilingual Information Extraction. In: Studies in Health Technology and Informatics. Medical and Care Compunetics 3, Vol. 121, 2006, p. 183-190, ISBN 978-1-58603-620-1.

[26] Moens, M.F.: Information Extraction: Algorithms and Prospects in a Retrieval Context. *Springer*, 2006. ISBN 1-4020-4987-0.

[27] Peshkin, L., Pfeifer, A.: Bayesian Information Extraction Network. In: Proc. *Intl. Joint Conference on Artificial Intelligence*, 2003.

[28] Praks, P., Dvorský, J., Snášel, V.: Latent semantic indexing for image retrieval systems. In: Proc. *SIAM Conference on Applied Linear Algebra*, Williamsburg, 2003.

[29] Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. In: Proc. *IEEE* 1989.

[30] Schroeder, I.: A case study in part-of-speech tagging using the ICOPOST toolkit. Technical report. University of Hamburg, Computer Science Department, 2002.

[31] Srovnání výsledku extrakčních nástrojů na úloze oznámení o seminářích, ITC-irst, http://tcc.itc.it/research/textec/tools-resources/learningpinocchio/CMU.

[32] Takeuchi, K., Collier, N.: Use of Support Vector Machines in Extended Named Entity Recognition. In: Proc. *CoNLL* 2002.

[33] Toman, J.: Srovnání přístupů extrakce užitečné informace z webu. In Proc. *Znalosti* 2008.

[34] Witten, I. H., Frank, E.: Generating Accurate Rule Sets Without Global Optimization. In: Proc. The 15th Intl. Conference on Machine Learning, Morgan Kaufmann, CA, 1998.

[35] Yates, A., Etzioni, O.: Unsupervised Resolution of Objects and Relations on the Web. In: Proc. *HLT* 2007.

[36] Žabokrtský, Z., Ptáček, J., Pajas, P.: TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In Proc. *WMT* 2008.

[37] Žáková, M., Štěpánková, O., Mařík̆ová, T.: MedAT: Medical Resources Annotation Tool. *KEG Seminar*, VŠE Praha, 2005. *http://keg.vse.cz/_slides/zakova.ppt*

Příloha A – Obsah disertační práce