СТАНОВИЩЕ

за присъждане на образователната и научна степен "доктор"
по научна специалност 01.01.12. - Информатика
на Борис Димитров Крайчев
докторант към катедра "Софтуерни технологии" на ФМИ, СУ "Климент Охридски"
на тема: "Извличане и анализ на мнения и чувства от текст от онлайн източници"
научен ръководител: доц. д-р Иван Койчев

от доц. д-р Красимира Иванова,
Институт по математика и информатика, БАН

Това становище е написано и представено на основание на заповед No. P038-57/05.02.2014 на Ректора на СУ "Св.Кл. Охридски".

1 Общо описание на представените материали и документи

Дисертационният труд е в обем от 114 страници. Съдържа резюме; увод; обзор на предметната област; три глави, съдържащи основните научни резултати в различни изследователски направления; една глава, описваща създадения софтуерен продукт, практическото му приложение и внедрявания; заключение, описващо основните приноси и плановете за бъдещо развитие; списък от 5 публикации, където са изложени части от дисертационния труд, 1 изнесен доклад на научна сесия на ФМИ, 4 цитирания на една от статиите и 6 цитирания на софтуерния продукт. Библиографията съдържа 85 литературни източника.

Представеният автореферат от 39 страници коректно отразява основните моменти на дисертационния труд, съответства по форма и по съдържание на изискванията.

Предоставени са 5 публикации, свързани с темата на дисертацията, три от които на международни конференции и две на национална конференция и национален семинар.

2 Актуалност на дисертационната тема

В последните години анализът на мнения от текст или изразено отношение или чувства придобива все по-голяма популярност. Той се различава от класическата категоризация на текст по тема по признаци като: (1) класификацията на текста по теми обикновено силно зависи от областта, докато при анализа на изразено отношение задачата се свежда до намиране на общи категории, разделени на две крайни мнения или по скала на силата на отношение; (2) в определянето на темите обикновено отделните стойности на класовете са независими докато тук те по-скоро са взаимно изключващи се или градиращи. Оттук произлизат и разликите в алгоритмите, които обикновено се прилагат; (3) самото описание на структурата на анализирания текст при анализа на мнения е по-еднотипно и може да бъде прилагано в различни области. Тези различия, обаче, не правят задачата по-лесна, както изглежда на пръв поглед, а показват единствено, че решаването й има по-широк спектър на приложение, отколкото решаването на даден проблем за анализ на факти в текстове от конкретна предметна област. Трудността на проблема идва от факта, че самото дефиниране на ключовите характеристики и тяхната сила силно зависят от субективната експресия на автора, нюанси в начина му на изказване, подвеждащите случаи на използване на сарказъм и ирония, и т.н.

Задачата за анализ на мнения от интернет източници добива все по-засилен изследователски интерес през последните години. Този вид проучвания имат голямо значение при маркетингови, социологически и други проучвания и за планиране на бизнес развитието.

3 Познаване на проблема

От моя гледна точка обзорът на областта "анализ на чувства от текст", направен във втора глава, е добре структуриран, пълен и логически издържан. Разграничени са различните нива, на които може да се извършва анализа – класическите "дума", "фраза/изречение", "документ" и новите възможности, които представят електронните среди за разглеждане на група смислово свързани в по-общо цяло единици документи. Дадени са основните методи, използвани в различните типове изследвания.

Авторът показва добро познаване на съвременните методи и средства, с които разрешава отделни стъпки от комплексната задача по анализ на текст от HTML страници, а липсващите елементи за изграждане на софтуерната среда, както и новите алгоритми, предложени от него, реализира самостоятелно, довеждайки до пазарен продукт.

4 Оценка на основните приноси на дисертационния труд

Един от основните приноси в дисертацията е предлагането и реализирането на алгоритъм за съпоставяне на йерархични структури с линейна сложност при извличане на информация от HTML документи. Алгоритъмът е използван в работата при извличането на потребителски мнения. Резултатите са публикувани в трудовете на международната конференция "Web Intelligence, Mining and Semantics", 2012. Изложението на метода и обосновката за сложността на алгоритъма са представени в трета глава на дисертацията. Добрата идея в случая е прилагането на алгоритъм от друга област (проектиране на компилатори) в областта на извличане на информация от структуриран текст. Доказана е приложимостта на идеята чрез програмна реализация и провеждане и анализ на експерименти върху общодостъпно тестово множество (Open Directory Project).

Друг принос е предлагането на един общ алгоритъм за класификация на потребителски мнения чрез самообучение. Алгоритъмът е най-общо със следните стъпки: (1) ръчно се избира начален набор от емоционално заредени думи, разширен с техните синоними от WordNet; (2) с помощта на техники за извличане на характеристики с етикетни последователни правила се извличат би-грами (прилагателно-съществително), които по-точно отразяват емоционалния заряд на речта; (3) използвайки построените лексикони се изгражда числено изображение на текста, представящо неговия емоционален заряд; (4) чрез машинно самообучение системата се настройва за конкретната среда, в която се извършва анализът.

Разработен е метод за автоматична класификация на прилагателните имена в българския език по емоционални оси. Следва да се отбележи, че наличието на такива изследвания за английския или други езици не може да бъде използвано чрез пряк превод на термините на български език, доколкото всеки език притежава специфична изразност.

Стройно са формулирани изискванията, на които един софтуерен продукт за анализ на онлайн репутация би следвало да отговаря – необходими функционални модули, изисквания към интерфейса с потребителя, към реализацията, съпровождането и поддръжката. Реализиран е цялостен продукт, който отговаря на формулираните изисквания.

Експериментите са проведени чрез паралелна обработка на големи масиви от неструктурирана информация.

Като добро начинание отчитам и даденият речник на съответствието между английската и българската терминология.

5 Бележки

Усеща се лека еклектика в методите и средствата, представени в трета, четвърта и пета глава – всяко от изследванията е само за себе си. Това, обаче, не е толкова рядко срещано в изследователските работи, свързани със софтуерни разработки – цялостният процес по зараждане на определена идея, въплъщението ѝ в програмна реализация и провеждането на експерименти е дълъг път, който чак накрая показва дали наистина е постигнат резултат, и една комплексна система придобива цялостен вид с течение на времето.

Описанието на предлаганите алгоритми и функции можеше да бъде подплатено с повече примери, показващи конкретно начина на работа на алгоритъм или конструиране на функция (напр. хеш-функцията) или определянето на полярни тегла на думите и др. – ако не в основния текст, то в приложения накрая.

Като цяло текстът е добре написан и форматиран. На няколко места, според мене, може да се преформатират фигурите с цел да не се оставя толкова празно място в текста (напр. фигура 3, фигура 17 и др.).

Вместо "Визуализация на връзката между термините с положителни и отрицателни емоции" на фигура 20 е повторена предишната визуализация от фигура 19. Самите три фигури (18, 19, 20) според мене се нуждаят от легенда или обяснение – аз лично не можах да се ориентирам откъде следваше твърдението "Чрез визуализация на основните атрибути може да се заключи, че отрицателните мнения имат по-висока сума на теглата на термини и фрази с негативни мнения и по-висок брой на думи, изразяващи противоречие". Може би и защото липсваше третата фигура.

В частта от изречението "Rankur е поставен под номер едно след други две приложения" (стр. 100) вероятно става дума за "пред други две приложения" (в посочения източник Rankur наистина е пред другите)?

Цитирането на литературата е чрез номер на източника, поставен в кръгли скоби – начин, който се използва в литературата. На стр. 29, 49 и 55, обаче, има изреждане на стъпки/условия/причини чрез използване на същата нотация. Желателно е да се намери друг начин за тяхното представяне.

Забелязани граматически грешки (неправилно членуване, липсващи или излишни предлози и др.) съм изпратила на автора за поправка.

6 Препоръки

В тази част излагам по-скоро неща, които в процеса на четене на работата възникваха като нови изследователски въпроси, подлежащи на бъдещо изследване.

В работата се разглеждат 4 варианта на оценки на потребителското мнение спрямо теглата на фразите с позитивен и негативен заряд и броя думи с противоречив заряд: score1: {posw}+{negw}; score2: {posw}+2*{negw}; score3: 2*{posw}+{negw}; score4: {posw}+{negw}-{contr}. Може би представлява интерес да се изследват оценки, вземащи пред вид и подредбата на групите думи с положителен и отрицателен заряд спрямо противоречивите. В ежедневието често употребяваме изречения от вида "(P), обаче (N)" – и в повечето случаи въздействието е с цел тотално унищожаване на положителния заряд в изречението.

При оценката на репутацията на обществени заведения обикновено потребителите изказват мнението си без да вземат пред вид изказванията на околните. В други случаи, обаче, в блог-пространството се развихрят канонади от спорове, които понякога изместват темата, от която са тръгнали. Бъдещо развитие на такъв тип изследвания може да разглежда не само всяко от мненията поотделно, но и в контекста на предишните.

По отношение на спам-съдържанието – посоченият начин за филтриране чрез проверка на правописа наистина би изхвърлил голяма част от негативните мнения, писани просто за заяждане или в афектно състояние. По отношение на прекалено положителните мнения би следвало да се прилага друг подход, доколкото някои компании, опитвайки се да заблудят потребителите сами "произвеждат" потребителски мнения в своя блог. Като индикатор за такъв тип поведение може би е интересно да се изследва разликата между представянето на компанията в конкретния сайт и мнения в други "независими" източници.

Горните коментари по-скоро още веднъж показват актуалността на тематиката и огромното поле за бъдеща работа в това направление.

Препоръката ми към дисертанта е да продължава в развитието на тематиката, доколкото добрите резултати от използването на програмния продукт Ранкър показват полезността на връзката между науката и бизнеса.

7 Заключение

През годините непряко съм запозната с разработваната тема от следенето на конференциите, на които са представени посочените от дисертанта публикации, и съм с отлични впечатления от работата. От моя гледна точка представеният труд напълно съответства по форма и съдържание на изискванията за дисертация за получаване на образователната и научна степен "доктор" по специалността "Информатика". Отчитайки и представените доказателства за реалното практическо използване на софтуерния продукт Ранкър убедено препоръчвам на членовете на уважаемото жури да гласуват за присъждане на тази степен на Борис Димитров Крайчев.

25.03.2014 Подпис: