Прилагане на количествени оценки при информационно търсене на документи

Б Ъ Л Г А Р С К А А К А Д Е М И Я Н А Н А У К И Т Е
ИНСТИТУТ ПО ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Светломир Минков Станчев

ДИСЕРТАЦИЯ
за придобиване на образователната и научна степен „доктор“
по специалност 02.21.10 Приложение на принципите и методите на кибернетиката в различни области на науката

Научен ръководител: проф. дтн Тодор Стоилов

- Начало на докторантурата 01.01.2009
- Край на докторантурата 31.12.2012

- Взети изпити:
  - Базов специализиран предмет
  - Интернет технологии за управление
  - Проектиране на интернет приложения
  - Компютърни умения - Java
  - Английски език

Публикации

- **Семантични мрежи и тяхното приложение в изграждане на уеб приложения, сп. Автоматика и информатика брой 1 Съюз по автоматика и информатика, 2012 28-33 стр.**
- **Модели информационно търсене, Доклад МЕЖДУНАРОДНА КОНФЕРЕНЦИЯ АВТОМАТИКА И ИНФОРМАТИКА’12 3 – 5 октомври 2012 г. София, България Съюз по автоматика и информатика.**

Основни понятия

Информационно търсене: Търсене на документи със символна неструктурирана информация в мрежата, които отговарят на информационно запитване на потребител.

Семантична зависимост между понятия. Семантичната зависимост е свързана с общите черти, които споделят тези понятия. Колкото повече общи черти споделят те, толкова зависимостта между тях е по-силна. Тези общи черти могат да се определят чрез включването им в онтологии, в които общи черти са общите понятия на по-високо ниво в семантичната мрежа

Статистическото езиково моделиране прилага количествени закономерности в оценката на смисловото съдържание на понятията и техните връзки между тях. Такива количествени оценки може да се използват за подобряване на работата на различни информационни приложения, обработващи информация на естествени езици

Информационно търсене

Модели информационно търсене

- Булев модел
- Латентно семантично индексиране
- Вероятностен модел
- Информационно търсене чрез статистически езикови модели

При всички тези модели на информационно търсене се предполага че понятията се срещат в информационното съдържание независимо едно от друго.

Проблеми при информационното търсене

Общ недостатък на прилаганите модели за информационно търсене е, че не се отчита семантичната зависимост между понятията(думите), които се срещат в информационното съдържание.

Не се взема предвид, че срещането на едно понятие може да повлияе на срещането на други понятия в дадено информационно съдържание, с които това понятие е в семантична зависимост.

Например ако се среща думата “море” вероятността да се срещнат и думите “кораб”, “остров”, “плаж” е много по-висока отколкото да се срещнат думи като “пустиня”, “камила”, “оазис”.

Проблеми при информационното търсене

Традиционните търсещи машини връщат огромен брой резултати от информационно търсене, като в тези резултати се съдържат както документи с високо семантично подобие на потребителската заявка, така и документи, които са включени защото съдържат ключова дума от заявката, но употребена в различен контекст.

Традиционните алгоритми за информационно търсене се базират на механично броене на ключови думи, без да отчитат тяхното информационно съдържание и контекста, в който са употребени тези ключови думи.

Предлагано решение

Семантичната зависимост между понятията може да бъде количествено формализирана чрез използване на методи, приложени върху онтологии, в които се срещат търсените понятия.

Тези методи се реализират чрез алгоритми за изчисляване на дължината на пътя между понятията в семантичния граф, върху който са базирани тези онтологии

На базата на измерената семантична зависимост между понятията може да бъде реализиран алгоритъм за информационно търсене с резултат по-малко на брой документи, но в по-голяма степен съответстващи на потребителската заявка за информационно търсене.

Предлагано решение

Ще бъде разработен алгоритъм за информационно търсене с използването на семантични езикови модели.

Ще бъде използвана количествена оценка на семантичното разстояние, за да бъдат проектирани и експериментирани алгоритми за информационно търсене в неструктурирани масиви от символна информация.

Целта е създаване на семантичен езиков модел, който е аналог на статистическия езиков модел, като разликата между тях е промененото вероятностно разпределение в модела с добавена семантична оценка на вероятността за срещаните понятия.

Чрез този семантичен езиков модел ще бъде измерено реалното вероятностно разпределение на понятията в този семантичен езиков модел, което ще позволи да се намерят документи със семантично съдържание, по близко до това на запитването.

Статистически езикови модели

Статистическото езиково моделиране може да се дефинира като функция на вероятностно разпределение на всички възможни езикови конструкции в един естествен език.

Задачата за построяване на статистически езиков модел отговаря на следния въпрос: Колко пъти се среща n-тата дума като се имат предвид предходните n – 1 думи.

Типове статистически езикови модели:

- Unigram модели n = 1
- Bigram модели n= 2
- N-gram модели

Семантична зависимост между понятия

Съществуват множество алгоритми за измерване на семантичната зависимост между понятия, които са базирани на измерването на дължината на пътя между съответните върхове в семантичния граф на избрана онтология.

Методът на броене на дъгите се базира на предположението, че колкото по-малко на брой дъги има между две понятия (върхове), толкова по-силна е тяхната семантична зависимост.

Методът за броене на дъгите има множество модификации, които подобряват неговата точност, като например въвеждане на тегла на дъгите, които са част от пътя между два върха или отчитане на дълбочината на върховете (разстоянието до корена).

WordNet

Като еталонна онтология е избрана и използвана WordNet.

Wordnet е голяма лексикална база данни, обхващаша думите от английския език.

Във WordNet съществителните, глаголите, прилагателните и наречията (за всяка част от речта е изградена отделна обособена мрежа) са групирани в синонимни групи, всяка от които изразява отделно понятие.

Синонимните групи са свързани помежду си с помощта на семантични и лексикални отношения.

В резултат понятията са свързани със семантични отношения и така създадения граф може да се обхожда. Структурата на WordNet го прави полезен инструмент за компютърна лингвистика.

WordNet е създадена и се поддържа от Когнитивната Научна лаборатория на Университета Принстън под ръководството на професора по психология Джордж А. Милър.

Причини да бъде използвана WordNet

- WordNet е голяма база данни, съдържаща голяма част от понятията в естествения език
- Съществуват алгоритми за измерване на семантичното подобие в WordNet
- WordNet е със сравнително постоянна структура, която се разширява, но вече създадената структура не се променя, което дава възможност за съхранение на изчислените разстояния между понятията

Предлаган алгоритъм за информационно търсене

1. Информационно търсене чрез традиционна търсеща машина
2. Подготвителни стъпки за построяване на семантичен езиков модел на всеки документ, резултат от традиционното търсене.
3. Построяване на семантичен езиков модел на всеки документ от списъка с резултати
4. За всеки документ намиране на вероятността документа и заявката да имат един и същ семантичен езиков модел.
5. Класиране на документите в съответствие с техните вероятности от стъпка 4

Информационно търсене чрез традиционна търсеща машина

Ще бъде използвана традиционна търсеща машина, която връща голям брой резултати, сред които се съдържат както документи, които удовлетворяват потребителската заявка, така и много неподходящи.

Известно е например, че Google намира почти всички документи, които съдържат ключовите думи от потребителската заявка.

В следващите стъпки на алгоритъма тези резултати ще бъдат филтрирани и класирани на базата на тяхното семантично подобие на потребителската заявка за информационно търсене

Подготвителни стъпки за построяване на семантичен езиков модел

Преди да бъдат построени семантични езикови модели на документите е необходимо да бъдат направени някои подготвителни стъпки.

1. Премахване на форматирането (например HTML тагове, препинателни знаци, и др.).
2. Преобразуване на думите в основната им граматическа форма
3. Премахване на често срещаните думи(местоименията на, по, във и т.н.)

Построяване на семантичен езиков модел

Включва следните етапи:

- Определяне на семантичната зависимост на всяко понятие в съдържанието на документа с всички останали чрез WordNet
- Намиране на вероятностното разпределение на понятията в текста на базата на тяхната семантична зависимост

Измерване на подобието между моделите на заявката и всеки документ

Ще бъде използван методът за подобие на заявката аналогично на информационното търсене чрез статистически езикови модели.

При този метод потребителската заявка се разглежда като съдържание, което може да се среща в документ. Ако то се среща, този документ се нарича идеален документ.

Задачата за измерване на подобието се свежда до измерване за всеки документ каква е вероятността той да бъде идеалния документ.

Ще бъде направено изследване и какви други методи могат да бъдат разработени и използвани за повишаване на точността и ефективността на измерването на семантичното подобие.

Планирани етапи

1. Допълнителни проучвания в областта на статистическите езикови модели и алгоритмите за намиране на семантичната зависимост между понятия.
2. Имплементиране на описания алгоритъм за информационно търсене чрез семантични езикови модели.
3. Експериментални резултати

Време за завършване на дисертационния труд приблизително 1 година

Благодаря за вниманието

Въпроси ?