

ЗВ'ЯЗАНИЙ ТА СЕМАНТИЧНИЙ ПОШУК: ОГЛЯД ПРАЦЬ

Цілі даної праці: пошук та аналіз літературних джерел, що містять дослідження можливостей семантичного та пов'язаного пошуку даних, а також пошук прикладних програмних засобів для роботи з ним, що могли б бути використані у подальших дослідженнях. Завдання дослідження: виявлення та реалізація напрямків використання існуючих семантичних засобів, зокрема, тезаурусу WordNet, в задачах обробки і упорядкування текстів. Предмет дослідження – алгоритми інформаційної обробки та/або теоретичні обґрунтування, описані у різноманітних джерелах. Методи дослідження – в статті використані спеціальні методи дослідження процесів і явищ: методи системного аналізу та синтезу. При роботі із теоретичним матеріалом, періодичними даними віднайшли своє застосування методи класифікації, порівняльного аналізу, угруповань.

Ключові слова: семантичний пошук, онтологія, зв'язані дані, інформаційний аналіз.

Вступ

У зв'язку з постійно зростаючою кількістю сайтів зростає потреба в ретельному аналізі контенту Інтернет-документів для того, щоб звести можливість отримання нерелевантних результатів до мінімуму. Технології семантичної павутини надають можливості для вирішення цієї проблеми.

Автором поставлена мета створення алгоритму пошуку зв'язаних даних у великому просторі невпорядкованих даних, який би допомагав користувачу отримати максимум інформації не тільки з шуканого предмета, але й з основ суміжних (наприклад, при пошуку інформації про садові інструменти також подаються короткі посилання на пошук мастила, інформації про ґрунтові добрива, і т.п.). Областю застосування алгоритму плануються зробити соцмережі як джерело постійно поновлюваного, хаотичного за змістом текстового джерела даних.

Метою даної роботи є пошук та аналіз літературних джерел, які б дали початкове уявлення про предмет роботи та дали напрямки щодо його удосконалення.

Виклад основного матеріалу

Було вирішено розділити огляд літератури на дві частини – теоретичну, в якій мали б бути описані суто лінгвістичні та семантичні аспекти досліджуваної проблеми, та прикладну, де містився би опис пошуку шляхів реалізації завдання.

1. Теоретична частина

На початку роботи автор намагався з'ясувати, що собою являє пошук інформації з точки зору лінгвістики і який чином це відбувається. У результаті пошуків було зроблено висновок, що основним і найбільш поширеним видом пошуку є семантичний, який великою мірою здійснюється за допомогою правильно підібраних онтологій.

Семантичним аналізом називається етап в послідовності дій алгоритму автоматичного розуміння текстів, що полягає у виділенні семантичних відносин, формуванні семантичного уявлення текстів. Один з можливих варіантів представлення семантичного уявлення - структура, що складається з "текстових фактів" [1].

1.1. Загальний аналіз семантичного пошуку

Виходячи з цього тлумачення, автори праці [2] розглядають існуючі технології семантичного пошуку і визначають специфічні проблеми, що стосуються пошуку документів в семантичній павутині з використанням запитів природною мовою. Згідно з їхньою думкою, семантична павутина (Semantic Web) є розширенням традиційного Інтернету і націлена на спрощення пошуку і розподілу інформації. Дана технологія ґрунтується на елементах, побудованих з використанням стандартних мов онтологій, таких як OWL. Звичайні пошукові системи ґрунтуються на пошуку ключових термінів запиту в документі і не можуть використовувати його смислове значення для отримання результату, тому спільнота дослідників семантичної павутини запропонувало використовувати семантичні пошукові технології.

Документ семантичної павутини SWD (Semantic Web Document) можна розглядати як набір даних, контентом якого є або онтологія, або звичайний документ, розмічений певними тегами, взятими з онтології предметної області.

Надалі у цій праці автори пропонують розглянути існуючі технології семантичної павутини в контексті наступних проблем: автоматичне створення формального запиту (онтології запиту) і отримання колекцій документів, структура яких не відома заздалегідь (розподілені і семантично різномірні дані). Стверджується, що найпопулярнішим видом пошуку є булевий пошук, заснований на комбінаціях ключових слів, розділених операторами AND, OR, NOT. Крім цього, існують також нечіткий пошук, пошук за контекстом, предметно-орієнтований пошук та пошук у тезаурусі. Технологія ключового пошуку може бути основою для отримання SWD-документів шляхом зіставлення шуканих слів поняттям, які відповідають онтологічним елементам в SWD.

Автори праці розкрили основні положення загального алгоритму побудови семантичного пошуку, який об'єднує декілька технологій і дозволяє створити мета-рушій для фільтрації SWD-документів у системі Swoogle - системі, що заснована на індексуванні пошуковим роботом SWD-документів с RDF (S) -, DAML- або OWL-синтаксисом [3]. Даний алгоритм може бути корисним, оскільки він дозволяє перетворювати запити природною мовою у формальні запити (онтології запиту). Для цього використовується словник WordNet, який використовує механізм автоматичного відображення на смислові значення для вирішення неоднозначності термінів запиту. Крім цього, алгоритм дозволяє проводити ранжування отриманих SWD-документів, ґрунтуючись на тому, наскільки добре вони відповідають онтології запиту: це визначається числом відповідностей між онтологією запиту і SWD-документом. Чим більше відповідностей між онтологією запиту і SWD-документом, тим вище позиція SWD-документа в фінальній видачі. Фактично набір SWD-документів, який бере участь в ранжируючому алгоритмі, є списком докумен-

тів, отриманих шляхом передачі запиту в довільній формі в Swoogle. Таким чином, ранжування може бути представлено як фільтруюча обробка SWD-документів, що повертаються ключовим пошуком.

Даний алгоритм складається з таких стадій:

1) *переформулювання запиту та усунення його неоднозначності;*

На цьому кроці для кожного терміна вільного запиту усувається неоднозначність, оцінюється передбачуване значення, яке визначається смисловим значенням WordNet. Для відображення терміна запиту на відповідне значення ми оцінюємо семантичну схожість кожного терміна з безліччю смислових значень WordNet. Набір значень WordNet можна вибрати за допомогою лексичної відповідності терміна з контентом поняття WordNet. Алгоритм бере до уваги околицю V_t кожного терміна запиту t .

2) *застосування алгоритму векторної моделі (Vector Space Model – VSM) для відображення терміну запиту на значення WordNet;*

Термін запиту t представляється як документ («набір слів», представлення). Оскільки термін t пов'язаний з усіма термінами в його околиці (V_t), документ, що представляє t , містить всі терміни, що зустрічаються в V_t . Крім того, кожне значення WordNet S_1, S_2, \dots, S_m , що представляє m можливих смислових значень терміна t , представлено у вигляді документа.

Ваговий вектор виду (w_1, w_2, \dots, w_N) , де $w_i, i = 1, \dots, T$, - це tf-idf значення відповідного терміна i , обраного з пов'язаних значень WordNet або термінів запиту в околиці V_t . Значення w_i терміна i підраховується наступним чином:

$$w_i = tf_i \times idf_i, \quad idf_i = \log_2 N/n_i,$$

де tf_i (term frequency, частота терма) - число входжень терміна i в окремому документі; idf_i - інверсія числа документів, що містять слово i ; N - загальне число документів; n_i - число документів, які містять терм i принаймні один раз.

У разі використання WordNet передбачуване значення терміна t визначається за допомогою VSM з усіх можливих значень (S_1, S_2, \dots, S_m) відповідного поняття WordNet.

Відображення терміна запиту на документ (значення в разі WordNet або набір назв, атрибутів і описів поняття в разі пов'язаної онтології) підраховується шляхом вимірювання відстані між вектором запиту q і вектором кожного документа. Результатом є упорядкований список документів. Документ з найвищим значенням косинуса подібності (cosine similarity) являє собою передбачуване значення терміна t . Він підраховується наступним чином:

$$Sim(w_i, w_j) = \frac{\sum_{k=1}^T w_{ik} w_{jk}}{\sum_{k=1}^T (w_{ik})^2 \times \sum_{k=1}^T (w_{jk})^2}.$$

3) *створення онтології запиту;*

Маючи відображення термінів на значення WordNet, ми можемо створити триплети, що включають поняття і відносини між ними. Залежно від того, що використовувалося для визначення передбачуваних значень термі-

нів запиту (пов'язана онтологія або словник типу WordNet), будуть застосовуватися різні правила для створення онтології запиту.

4) отримання і ранжування SWD-документів.

На цьому кроці алгоритм перевпорядковує SWD-документи, отримані з пошукової системи Swoogle. Повторне ранжування засноване на відкритій семантиці термінів запиту, переформульованих в онтологію запиту, і семантиці отриманих SWD-документів.

Автори акцентують увагу на тому, що у алгоритму є кілька проблемних критичних точок, а саме: продуктивність (швидкість виконання запиту) та точність SWSS-системи - для запиту SWD-документів в реальних системах технології і реалізації, що використовуються, повинні бути протестовані і оцінені в контексті точності і повноти одержуваного результату.

1.2. Побудова семантичної мережі шляхом видобуття знань

Інший та складніший спосіб побудови семантичної мережі пропонують Найханова Л.В. та ін. у [4]. Дана робота присвячена вирішенню завдання побудови семантичної мережі номенклатури предметної області. Вважається, що перед початком роботи була виконана попередня лінгвістична обробка наукового тексту Θ , в результаті якої були отримані безліч лексем, векторів, стійких словосполучень, безліч векторів морфологічної інформації лексем і статистичної інформації про лексеми і про стійкі словосполучення. Також в початковій точці є безліч модифікованих графів залежностей $G = \{g_i \mid i = 1 \dots n, n - \text{кількість речень в тексті } \Theta\}$, в яких окремі лексеми об'єднані в словосполучення, і безліч графів семантичної околиці терміна $G^* = \{g_i^* \mid i = 1 \dots q, q - \text{кількість графів в тексті}\}$.

Метою описуваної праці є необхідність побудувати єдину семантичну мережу S на основі аналізу графів G і G^* .

Авторами роботи пропонується спочатку розглянути категоріальний апарат семантичної мережі, тобто способи і види відносин між словами - ієрархія (рід - вид), агрегації (ціле - частина), тотожності (термін - його синонім), і т.д., а також перерахувати і впорядкувати категорії концептуальних об'єктів мережі - вони включають «поняття», «дія», «стан», «подія», «величини». На основі конструкцій знаків концептуальних об'єктів визначені структури словникових статей, що дозволяють описувати конкретні об'єкти.

Після цього описується безпосередньо побудова семантичної мережі. Семантична мережа S є об'єднанням елементарних фрагментів Φ :

$$S = \bigcup_i \Phi_i$$

де Φ - це предметна область. Побудова семантичної мережі починається з побудови елементарного фрагмента, який будемо називати стартовим фрагментом. Вибір терміна t_0 для побудови стартового фрагмента семантичної мережі пропонується здійснювати на основі аналізу безлічі графів семантичної околиці терміна G_i^* . Для обраного терміна t_0 будується елементарний фрагмент Φ_0 .

Після цього для терміна вираховуються за особливими складовими величини знаків, відповідні категоріям об'єктів - тобто визначаються безлічі дефініцій терміна, безлічі станів, в якому знаходиться термін (наприклад, як-

що термін - "вода", то безліч включає "кипить", "вариться", "замерзає"), і т.д. Після цього проводиться об'єднання і аналіз графів залежностей.

На думку авторів, підхід, що описується в їх роботі, може стати основою технології створення загальної онтології з часткових. Крім того, даний підхід може бути використаний при побудові пошукового образу текстового документа.

2. Практична частина

Розглянувши огляд праць, в яких описуються теоретичні засади побудови семантичної мережі, пропоную перейти до розгляду робіт, в яких автори практично тією чи іншою мірою реалізують пошук залежних даних.

2.1. Побудова ранжованої семантичної структури набору документів

Один з найпростіших способів побудови семантичної павутини пропонується в [5], а саме – з колекції текстових документів. Основним етапом, на думку автора, має бути ранжування слів-кандидатів з використанням статистичної інформації – ранг слова залежить від величини «локального» та «глобального» TF-IDF (Вага (значимість) слова, що пропорційна кількості вживань цього слова у документі, і обернено пропорційна частоті вживання слова у інших документах колекції). «Локальним» TF-IDF вважається такий спосіб ранжування, у якому ключові слова займають перші x %:

$$rank_i = \frac{n_i}{\sum_{j=0}^N n_j} \log \left(\frac{|DOC|}{|\{doc : d_i \in doc\}|} \right), \quad D_{key} = \left\{ d_i : \frac{x}{100} |D| \leq rank_i \right\}$$

а «глобальним» - такий, у якому ключові слова – це об'єднання перших x %, але не більш ніж для u слів.

$$rank_{ij} = \frac{n_{ij}}{\sum_{j=0}^N n_{ij}} \log \left(\frac{|DOC|}{|\{doc : d_i \in doc\}|} \right), \quad D_{key} = \bigcup_{j=1}^{|DOC|} \left\{ d_i : \min \left\{ y, \frac{x}{100} |D_j| \right\} \leq rank_{ij} \right\}$$

Величини x та u підбираються емпірично.

2.2. Модифікація онтологічних мереж

Посилаючись, зокрема, на роботу [5], автори [6] визначають підходи і методи для реалізації коректного механізму оновлення онтологій на основі семантичних мереж. Стверджується, що завдання оновлення онтологічних знань нетривіальне і в основному може вирішуватися в напівавтоматичному режимі, спільно з експертом. Відповідно, метою праці ставиться аналіз методів і алгоритмів, що дозволяють оновлювати онтологію по семантичній мережі в напівавтоматичному режимі.

Для реалізації конкретних питань на першому етапі необхідно вибрати найбільш ефективний і в той же час швидкий спосіб побудови семантичної мережі. З безлічі існуючих методів і був обраний метод створення семантичної мережі з колекції текстових документів певної предметної області [5].

Для отримання можливості поновлення онтологій необхідно на початку створити корпус концептів, для цього слід провести виділення ключових слів, ключових словосполучень і групування словосполучень. У свою чергу виділення ключових слів складається з: нормалізації, токенизації, лематизації. Далі слідує фільтрація, тобто видалення всього, крім іменників і прикметни-

ків. Нарешті, слова-кандидати ранжуються з використанням статистичної інформації.

Виділення ключових словосполучень також ділиться на окремі кроки. Слід видобути вільні словосполучення, потім згрупувати словосполучення-кандидати шляхом пошуку найбільших спільних підстрок і відранжувати їх.

Після реалізації всіх кроків, описаних вище, отримаємо семантичну мережу.

Наступним кроком, за словами авторів, є безпосередньо оновлення онтології, яке так само складається з декількох етапів і при якому вирішення конфлікту імен проводиться експертом. В результаті формуються два списки: список співпадаючих імен концептів і список концептів, які не збігаються, що містяться в семантичній мережі і доповнюють базову онтологію.

Нарешті, кожен елемент другого списку обходиться з визначенням його прямого батька, потім, при наявності подібного з певним батьком класу в вихідній онтології, новий концепт безпосередньо приєднується до онтології.

2.3. Алгоритм пошуку пов'язаних даних у неструктурованій мережі

Важливою вихідною точкою для автора цієї статті стала праця А. М. Глибовця [7], в якій автор подає ідею контекстного пошуку залежних даних, визначає основні поняття та подає спрощене керівництво до дії. Також наведено початковий аналіз інформаційних ресурсів для пошуку пов'язаних даних. Автор подає початковий алгоритм пошуку зв'язків між даними у неструктурованому наборі даних (пошуковій видачі) та спосіб його реалізації, що заснований на поєднанні латентно-семантичного та частотного аналізів. Основна особливість алгоритму полягає в тому, що підраховується кількість та якість слів, що знаходяться у наперед заданому околі ключового слова. В кінці цієї статті надається приписка: «Шляхами удосконалення алгоритму багачиться покращення роботи з соціальними мережами ...».

2.4. Пошук та усунення двозначності в даних семантичного пошуку

В кінці пропоную розглянути роботу [8], в якій піднімається проблема неповного (нечіткого) дублювання даних в реляційних БД і способи її вирішення. Можна виділити два основних типи дублювання атрибутів: такі, що мають жорстко задану структуру (формат) змісту і не мають такої, тобто неповно структуровані. Автори зосередилися саме на другому типі як такому, що не має однозначного вирішення.

Для вирішення вищезгаданих проблем був розроблений підхід, цілями якого було: введення в розгляд об'єктивних чисельних характеристик якості інформації за критерієм її дублювання і опис ключових етапів процесу щодо забезпечення якості на основі цих введених характеристик.

Автори стверджують, що в загальному випадку при порівнянні окремих слів і простих словосполучень хороші результати дає застосування методу n-грам в поєднанні з коефіцієнтом Дайса. Коефіцієнт обчислюється таким чином [9]:

$$dice(X, Y) = 2 * (|n\text{-grams}(X) \cap n\text{-grams}(Y)|) / (|n\text{-grams}(X)| + |n\text{-grams}(Y)|).$$

Тут X , Y - рядки; $n\text{-grams}(X)$ - безліч грам рядка X ; $|N\text{-grams}(X)|$ - потужність множини.

Але при порівнянні речень і фраз починають проявлятися смислові помилки. Наприклад, якщо інформація про контактну особу порівнюється у вигляді рядків, то загальний коефіцієнт схожості може бути високим, тоді як за деякими смисловими одиницями фрази схожість низька.

З огляду на цю особливість, автори пропонують ввести інтегральну оцінку схожості для унікальних ідентифікаторів записів в БД. Зважена схожість $\text{sim}w$ двох ідентифікаторів I_{k1} і I_{k2} обчислюється за формулою:

$$\text{sim}w(I_{k1}, I_{k2}) = \sum_{i=1}^N \text{dice}(A_{i1}, A_{i2})(w_{i1} / N).$$

де A_{k1}, \dots, A_{kn} - атрибути ідентифікатора I_k , а w_{j1}, \dots, w_{jn} , - смислові ваги j -го атрибуту, $w_j \in (0..1]$.

Далі описується порядок дій, необхідний для пошуку дублікатів відповідно до наведеної формули, експериментальним шляхом для тестового масиву даних визначається нижній поріг зваженої схожості, за яким кількість помилок розпізнавання дублікатів стає неприйнятною. Це значення пропонується назвати порогом автоматичної обробки і позначити як Π_a .

Переходячи до практичного застосування методу, завдання виявлення і усунення дублікатів в інформаційній системі розбивають на три етапи:

1) первинне виявлення дублікатів на рівні введення інформації користувачами та визначення їх відхилення, якщо $\text{sim}w > \Pi_a$;

2) виявлення дублікатів шляхом порівняння і аналізу вже введених даних відповідно до заданого Π_a і автоматичне видалення дублюючої інформації, якщо $\text{sim}w > \Pi_a$;

3) аналіз і обробка людиною результатів п. 2, які не можуть бути оброблені автоматично, тобто $\text{sim}w < \Pi_a$.

Враховуючи, що будь-яка соцмережа технічно являє собою базу даних, дана праця може видатися корисною при ручній перевірці та очистці результатів власної роботи автора поточної праці.

Внесок автора

Виходячи з пропозиції А.М. Глибовця в [7] про принципову можливість застосування розробленого ним алгоритму пошуку пов'язаних даних у соціальних мережах, мною було прийнято рішення провести адаптацію та пристосування алгоритму до соцмережі «ВКонтакте» та перевірка, чи можливо отримати кращі результати (відсоток релевантності), ніж при використанні вбудованого пошуку самого ВК.

Автором була написана програма імплементації алгоритму з [7] в соцмережі на мові С#, яка здійснювала пошук та фільтрацію результатів. Програма показала прийнятні результати – близько 60% релевантних результатів пошуку. Незважаючи на таке відносно мале число, воно є допустимим тому, що у ВК є дуже багато шумових текстів, які, формально являючись результатами пошуку, не є такими (тобто тексти, створені спеціально для протягування аудиторії), тому при визначенні оцінки якості роботи програми автор вирішив їх не враховувати.

Наступним кроком я запланував використання результатів пошуку для створення мережі зв'язків між ними, щоб користувач міг здійснювати, зокрема, перехресний пошук, що ґрунтується на результатах попередніх пошуків. Для здійснення цього задуму потрібно створити т.зв. синонімічну таблицю, що є прототипом онтології. Це можна реалізувати способами, описаними в [6]. Надалі, застосовуючи алгоритми латентно-семантичного аналізу (LSA) у зв'язці з тезаурусами WordNet, як це вказано в [2], планувалося побудувати розлогу таблицю синонімів та пов'язаних слів (я вирішив обмежитись лише іменниками з огляду на їх відносну простоту сприйняття), що й було зроблено. Тепер я маю намір використати її для перехресного повторного пошуку.

Ранжування результатів поки що було вирішено не проводити з огляду на швидку динаміку оновлення контенту у ВК, але за потреби це можна зробити за допомогою описаних у [5] формул.

Висновки

У даній роботі було проведено аналіз літератури з теми пошуку пов'язаних даних в та поза семантичною мережею, даний їх короткий опис та характеристика. Описані їх сильні сторони в контексті використання для реалізації авторського алгоритму. Подано короткий опис авторського внеску.

Бібліографія

1 Леонтьева Н. Н. К теории автоматического понимания текста. Ч. 3. Семантический компонент. Локальный семантический анализ. — Изд. Моск. ун-та Москва, 2002. — С. 49.

2 Басипов А.А., Демич О.В. – Семантический поиск: проблемы и технологии. – Вестник АГТУ. Серия: управление, вычислительная техника и информатика. – №1, 2012. – С. 104–111.

3 Bernstein A., Kaufmann E., Fuchs N. Talking to the semantic web – a controlled english query interface for ontologies // AIS SIGSEMIS Bulletin. – 2005. – N 2. – P. 42–47.

4. Найханова Л. В., Аюшеева Н. Н., Хаптахаева Н. Б. – Построение семантической сети предметной области на основе извлечения знаний из научного текста – Известия высших учебных заведений. Поволжский регион. Технические науки – №4, 2007. – С. 51-61.

5. Паченко А.И. - Построение семантической сети из разнородных данных. – М.: МГТУ им. Баумана, 2010. – Режим доступа: http://it-claim.ru/Persons/Panchenko/presentation2010_sept_final.pdf

6. Вороной С.М. Обновление онтологий с помощью семантических сетей текстов на естественном языке / С.М. Вороной, А.С. Калинин, К.С. Охрименко // Информационные управляющие системы и компьютерный мониторинг. – Донецк : ДонНТУ, 2012. – С. 83-85.

7. Глибовець А.М. – Алгоритм пошуку зв'язків і залежностей між даними веб-сторінок. – Проблеми програмування: Науковий журнал. - №1, 2016.

8. Бураков В.В., Тарасов С.В. – Контекстно зависимый способ поиска нечетких дубликатов в реляционных базах данных – Информационно-управляющие системы – №2, 2015 – СПб: ГУАП, 2015. – С. 76-81.

9. Мазов Н. А. N-граммные методы обработки текстовой информации/ОИГГМ СО РАН. – Новосибирск, 1995. – 180 с.