УДК 001.8:004.7

Д. В. Ланде, С. В. Прищепа, Т. В. Синькова
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Автоматична побудова термінологічної мережі як моделі предметної області

Запропоновано алгоритм побудови термінологічних мереж — моделей предметних областей на основі репрезентативного набору тегів, отриманого в результаті зондування великої інформаційної мережі. За основу взято мережу понять, які відповідають тегам наукометричного сервісу Google Scholar Citations, вузли якої — поняття, марковані тегами, а ребра — деякі семантичні зв'язки між ними, обумовлені суміжними інтересами окремих авторів. Наведено правила автоматичної побудови на основі даної мережі бібліографічного списку релевантних публікацій. Запропонований підхід може бути застосовано, зокрема, до бібліографічних баз даних, в яких у явному вигляді виділені автори, і як теги — ключові слова, а також для багатьох напрямків науки.

***Ключові слова**: модель предметної області, Google Scholar Citations, бібліографія, зондування мережі, візуалізація мережі.*

Задача створення моделі предметної області

При вивченні та заглибленні в будь яку предметну область (ПрО) особливе значення має визначення переліку найважливіших понять цієї суміжної області, а також взаємозв'язку цих понять. Саме як модель ПрО сьогодні все частіше розглядають спеціальним чином сформовану мережу понять — онтологію. Побудова великої галузевої онтології, як правило, складна науково-практична проблема [1, 2]. Основа цього процесу — побудова термінологічної основи та визначення деяких семантичних зв'язків [3].

У даній роботі розглядається спроба формування моделі досить вузької предметної області — методів екстрагування подій із текстових масивів (Event Extraction). Як відомо, ці методи пов'язані з популярною концепцією глибинного аналізу текстів (Text Mining), більш поширеними технологіями глибинного аналізу даних (Data Mining), графів (Graph Mining) тощо. З іншого боку, поняття «екстрагування подій» пов'язане з близькими за значеннями поняттями «визначення подій»

(Event Detection), «виявлення фактів» (Fact Extraction), або «екстрагування інформації» (Information Extraction).

Саме визначенню місця поняття «екстрагування подій» серед інших близьких, побудові відповідної моделі предметної області методології присвячено цю статтю.

Зондування наукометричної мережі

Задача автоматичного створення онтологій вимагає враховувати знання, закладені фахівцями (науковцями, експертами) у деякі тексти. Як такі тексти можуть розглядатися спеціальні довідники, масиви документів [2], мережеві публікації тощо. Фактично, у рамках даної роботи розглядається метод побудови моделі ПрО, заснований на знаннях, закладених у спеціальних словниках, що формуються експертами — авторами наукових праць.

У роботі розглянуто підхід і алгоритми автоматизованого формування моделі ПрО шляхом зондування наукометричних інформаційних мереж. Під зондуванням великих інформаційних мереж будемо розуміти вибірку невеликого обсягу найважливішого змісту таких мереж, які з технологічних причин не підлягають повному скануванню.

Як велика інформаційна мережа розглядається мережа понять, які відповідають тегам наукометричного сервісу Google Scholar Citations, вузли якої — поняття, марковані тегами, а ребра — деякі семантичні зв'язки між ними, обумовлені суміжними інтересами окремих авторів.

На рис. 1 наведено фрагмент інтерфейсу сторінки результатів веб-сервісу Google Scholar Citations, що відповідає заданому заздалегідь для пошуку тегу Event Extraction (екстрагування понять). Цю сторінку можна отримати, вказавши у адресному рядку браузера вираз: http://scholar.google.com/citations?view_op=search_authors&mauthors=label:event_extraction

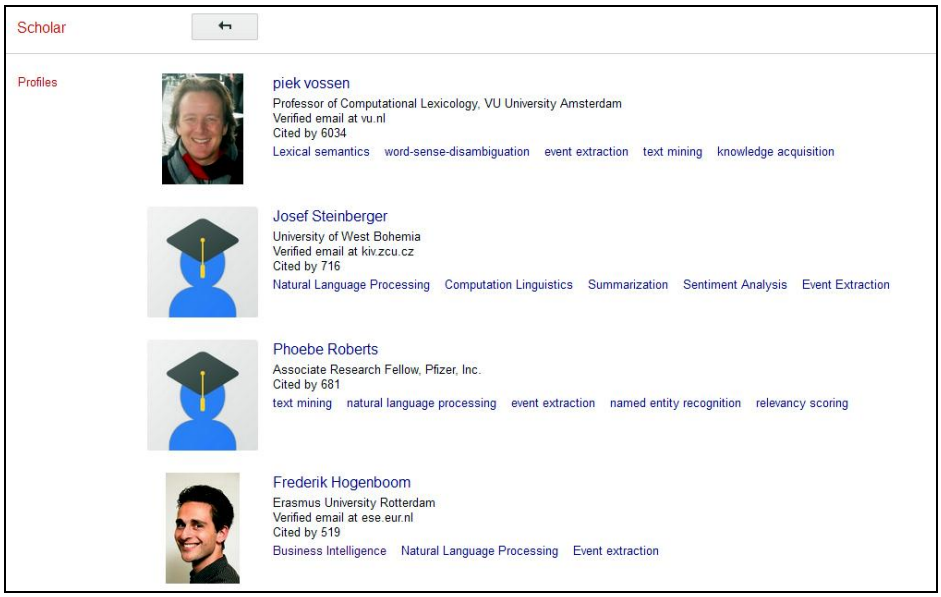


Рис. 1. Інтерфейс сторінки веб-сервісу Google Scholar Citations

Під тегом у рамках сервісу, що розглядається, розуміють позначення поняття, наукового напрямку, приписуваного вченому, яке відповідає його науковим інтересам. Після введення відповідного адресного рядка в інтерфейсі сервісу Google Scholar Citations у ранжируваному вигляді посторінково відображаються імена вчених, які позначили свої наукові інтереси заданим поняттям. Для кожного вченого вказано це поняття, а також інші поняття, що входять до сфери його наукових інтересів. Наприклад, для автора Piek Vossen визначені ще такі теги, як Lexical Semantics, Word-Sense-Disambiguation, Event Extraction, Text Mining, Knowledge Acquisition. Множина тегів-понять утворює мережу, похідну від біграфа «вчений-поняття» (тобто графа, множину вершин якого можна розбити на дві частини таким чином, що кожне його ребро з'єднує якусь вершину з однієї частини з якоюсь вершиною іншої частини). В рамках даної роботи як одну частину можна розглядати множину вчених, а іншу — множину тегів). Для моделі ПрО розглянемо похідну від цієї мережі, а саме мережу зв'язків окремих понять через авторів, яким ці поняття приписані. Очевидно, зв'язки такої «похідної» мережі можуть мати вагу, пропорційну кількості авторів, яким приписується відповідна пара понять.

Існують випадки, коли теги, зазначені окремими вченими, можуть належати до різних галузей науки, однак, попередньо проведені дослідження показують, що на невеликій, але досить репрезентативній вибірці (порядку сотні тегів), невелика частота нетематичних тегів забезпечує їхнє автоматичне «відсіювання» за деякім невеликим порогом.

Модель зондування мережі

Зондування модельної мережі здійснюється за принципом, подібним до того, що застосовується при пошуку інформаційних ресурсів у пірінгових мережах (від англ. Peer-to-peer, мережа — однорангова мережа, заснована на рівноправності учасників) [4–6], що складається з наступних кроків.

1. Вибирається певна кількість вузлів опорної (тієї, що зондується) мережі, які визначаються як базові для нової мережі, що є результатом зондування.

2. Для кожного з розглянутих вузлів опорної мережі визначаються суміжні з ним вузли («сусіди»), що додаються до створюваної мережі за результатами зондування.

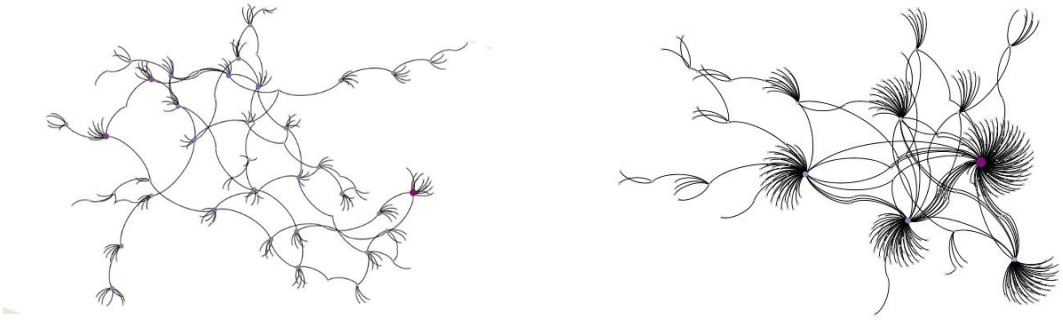
3. Від кожного поточного вузла опорної мережі здійснюється перехід до сусіднього вузла, що має найбільший ступінь, після чого здійснюється перехід до кроку 2.

4. Якщо має місце «зациклювання» (вибирається вузол, до якого вже було здійснено перехід за цим алгоритмом), відбувається перехід до наступного за ступенем сусіднього вузла. Якщо таких сусідніх вузлів не залишилося — здійснюється перехід до кроку 2.

5. Якщо перелік базових вузлів завершено, вважається, що мережа, яка відповідає результатам зондування, побудована.

Даний алгоритм перевірявся для двох найпоширеніших модельних мереж Erdős-Rényi (ER) і Barabási-Albert (BA) (рис. 2) [7, 8]. Відомо, що модель ER — це випадкова мережа, яка будується наступним чином: множина з N від початку

нез'єднаних вузлів попарно об'єднують з імовірністю p. У результаті створюється мережа приблизно з $p\frac{N(N-1)}{2}$ випадково вибраними зв'язками. Модель ВА — одна з декількох моделей мереж зі степеневим розподілом ступенів вузлів (так званих, безмасштабних мереж). Ця модель враховує як зростання мережі (динаміку), так і принцип переважного приєднання, який полягає в тому, що чим більше зв'язків має вузол, тим більш імовірно для нього створення нових зв'язків зі знову утвореними вузлами. Вузли з більшим ступенем мають велику ймовірність приєднання (створення нових зв'язків) до нових вузлів [7].



а) б)

Рис. 2. Приклад мереж, побудованих шляхом зондуванням модельних мереж:
а) Erdős-Rényi; б) Barabási-Albert

Із самого початку передбачалося, що мережі понять, що природним чином формуються учасниками мережевих сервісів, як і більшість інформаційних мереж мають властивість безмасштабності [8] (тобто близькі за структурою до мережі ВА). Якщо мережа така велика, як, наприклад Google Scholar Citations, на допомогу може прийти зондування, в результаті якого виконується побудова деякої нової мережі, що лише частково збігається з вихідною. Зазначимо, що результати будь-якого зондування не завжди вірно відображають природу великої досліджуваної мережі — вони багато в чому залежать саме від алгоритму процедури зондування, разом з тим, вони можуть служити базою для побудови гіпотез щодо структури первинної мережі.

Візуально якісні результати зондування мереж ER і ВА з близькими параметрами (1000 вузлів, близько 2000 зв'язків) наведені на рис. 2. Порівняння показує, що пов'язані області (гілки), що відповідають окремим поняттям у першому випадку достатньо довгі, а вузлів, по яким слідує маршрут зондування довший, ніж у другому, більш цікавому для дослідження реальних мереж, випадку. Зазначимо, що в рамках даного дослідження більш важливі саме якісні результати, вид пов'язаних ланцюжків, якими моделюються гілки понять. Аналіз результатів моделювання показав, що наведений алгоритм при зондуванні реальної мережі ймовірно буде швидко «зациклюватися» (і, відповідно, перериватися), що призведе до ще більшого скорочення гілок понять. Саме на підставі результатів цього якісного моделювання було зроблено висновок щодо можливості формування невеликих (осяжних) гілок пов'язаних тегів-понять сервісу Google Scholar Citations.

Алгоритм зондування мережі Google Scholar Citations

Наведений вище алгоритм, який застосовувався до модельних мереж, був адаптований до реальної мережі тегів сервісу Google Scholar Citations наступним чином.

1. Експертним шляхом визначається невеликий перелік базових тегів (ключових слів, відповідних найбільш важливим поняттями ПрО).

2. Вибирається тег із визначеного експертами переліку.

3. Відкриваються сторінки веб-сервісу, що відповідають цьому тегу (максимальна кількість таких сторінок параметрично обмежується заздалегідь).

4. До створюваної мережі додаються всі теги, що містяться на вибраних сторінках (сусідні теги).

5. Із сусідніх тегів вибирається той, на відповідні сторінки сервісу якого планується перейти для подальшого аналізу. Цей тег з найбільшим ступенем серед сусідніх тегів, який також задовольняє тематиці обраної ПрО і не входить до складу тих тегів, до сторінок яких вже було здійснено перехід.

6. Якщо такий тег обрано, то відбувається перехід до п. 3.

7. Якщо такого тегу не існує, але перелік базових тегів не завершено, то здійснюється перехід до наступного базового тегу з початкового переліку, тобто перехід до п. 2. У іншому випадку вважається, що мережа зондування побудована.

Відповідно до наведеного алгоритму процес зондування мережі, починаючи з певного вузла, припиняється при «зациклюванні», тобто коли відповідно до алгоритму відбувається перехід до вже пройденого тегу, а також при відхиленні від основної тематики сусідніх тегів, що залишилися.

Відхилення від основної тематики визначається експертами при автоматизованому зондуванні або з урахуванням лексичного складу тегів при повністю автоматичному скануванні. У разі автоматичного виконання алгоритму виконується обмеження за допомогою так званих «плюс-» і «стоп-словників» — наборів спеціальних шаблонів («плюс-» і «стоп-словник» у рамках даної роботи — набори шаблонів, підрядків, які повинні обов'язково входити або, відповідно, не входити в рядки, відповідні тегам). При цьому саме «зациклювання» є ознакою переходу до наступного базового тегу або завершення процесу зондування.

Формування стартового переліку вузлів-понять і правил відбору «кінцевих» вузлів виконується експертами з ПрО. Для побудови моделі ПрО (в розглянутому прикладі для області екстрагування подій із текстів) експертним шляхом були визначені базові теги англійською мовою:

Event Extraction;
Information Extraction;
Event Detection;
Entity Extraction;
Fact Extraction.

У якості «плюс-словника» використовувався такий набір шаблонів:

mining;
event;
extraction;
entity;

fact;
monitoring;
detection.

У розглянутому окремому випадку «стоп-словнику» відповідав набір шаблонів (зокрема, виключалася медико-біологічна тематика, туризм тощо):

bio;
medic;
health;
tourism;
venture;
portfolio;
comparative.

При наведених вище трьох базових тегах, за результатами зондування мережі Google Scholar Citations загальна кількість вузлів-тегів, які були екстраговані за алгоритмом, склала 148, у той час, як кількість нетермінальних вузлів — лише 14.

На рис. 3 наведено приклад центрального фрагменту мережі понять ПрО, побудованої відповідно до наведеного алгоритму за вказаними базовими тегами.

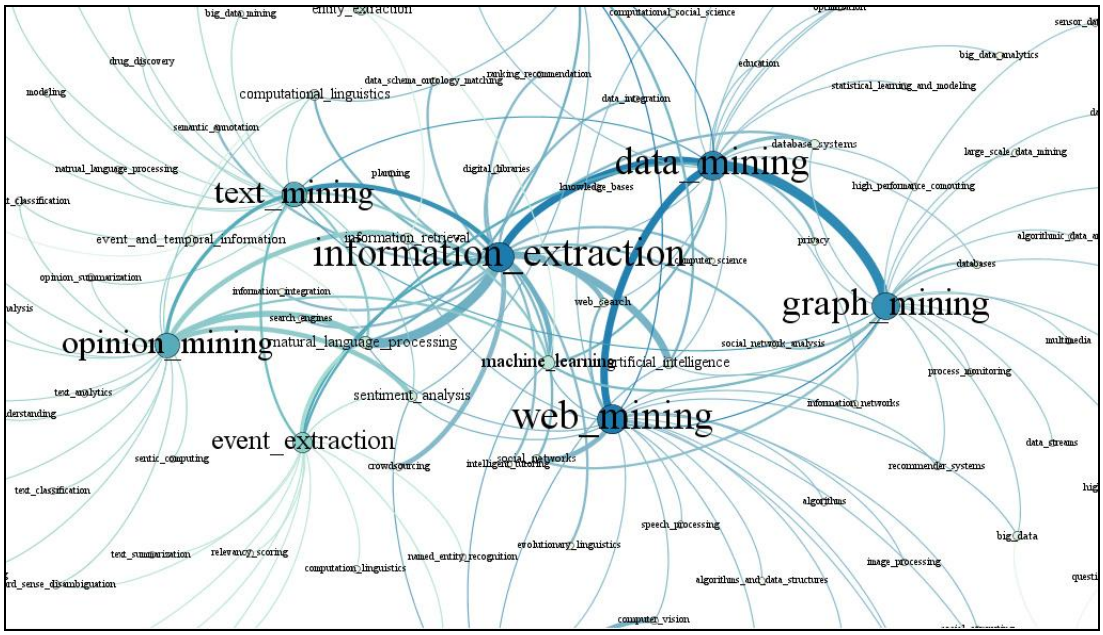


Рис. 3. Центральний фрагмент мережі понять

Як можна побачити, у цій мережі відбувся «зсув понять» (Topic Drift), тобто тег Event Extraction займає в ній далеко не центральне місце, яке природно займають такі більш поширені поняття, як Information Extraction, Data Mining, Web_mining. Разом з тим, саме місце поняття, що досліджується у цій мережі, є необхідною якісною властивістю даної ПрО.

Побудова бібліографічних списків

На підставі побудованої у результаті зондування сервісу Google Scholar Citations мережі і можливостей даного сервісу реалізовано алгоритми автоматичного складання бібліографічних списків, що відповідають найбільш цитованим роботам у вибраній області за вказані проміжки часу.

Правила побудови бібліографічних посилань досить прості і охоплюють наступні кроки.

1. Послідовно вибирається заздалегідь задана кількість найбільш вагомих вузлів побудованої мережі.

2. У відповідності з цими вузлами-тегами формуються запити до сервісу Google Scholar Citations і відбираються найбільш цитовані автори за вказаний період часу.

3. Відкриваються сторінки вибраних авторів, на яких наводяться посилання на публікації, із зазначенням їхнього цитування.

4. Серед найбільш цитованих робіт автора вибираються публікації, заголовки яких відповідають «плюс-» і «стоп-словникам», приклади яких наведені вище.

5. У разі необхідності, з відібраних публікацій вибираються тільки ті, які містять посилання на повні тексти у форматі PDF.

На рис. 4 наведено приклад автоматично сформованого бібліографічного списку з посиланнями на PDF-файли публікацій.

Bibliography

1. *AnHai Doan.* Declarative information extraction using datalog with embedded extraction predicates. W Shen, AH Doan, JF Naughton, R Ramakrishnan. Proceedings of the 33rd international conference on Very large data bases..., 2007.: 152
2. *AnHai Doan.* Source-aware entity matching: a compositional approach. W Shen, P DeRose, L Vu, AH Doan, R Ramakrishnan. Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on..., 2007.: 161
3. *Andrew McCallum.* A comparison of event models for naive bayes text classification. A McCallum, K Nigam. AAAI-98 workshop on learning for text categorization 752, 41-48., 1998.: 2576
4. *Andrew McCallum.* Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. A McCallum, W Li. Proceedings of the seventh conference on Natural language learning at HLT..., 2003.: 704
5. *Andrew McCallum.* Learning hidden Markov model structure for information extraction. K Seymore, A McCallum, R Rosenfeld. AAAI-99 Workshop on Machine Learning for Information Extraction, 37-42., 1999.: 486
6. *C Lee Giles.* Automatic document metadata extraction using support vector machines. H Han, CL Giles, E Manavoglu, H Zha, Z Zhang, E Fox. Digital Libraries, 2003. Proceedings. 2003 Joint Conference on, 37-48., 2003.: 335
7. *Daniel Weld.* Web-scale information extraction in knowitall:(preliminary results). O Etzioni, M Cafarella, D Downey, S Kok, AM Popescu, T Shaked,... Proceedings of the 13th international conference on World Wide Web, 100-110., 2004.: 721
8. *Daniel Weld.* Open information extraction from the web. O Etzioni, M Banko, S Soderland, DS Weld. Communications of the ACM 51 (12), 68-74., 2008.: 315
9. *Daniel Weld.* Unsupervised named-entity extraction from the web: An experimental study. O Etzioni, M Cafarella, D Downey, AM Popescu, T Shaked, S Soderland,... Artificial intelligence 165 (1), 91-134., 2005.: 891

Рис. 4. Фрагмент бібліографічного списку з посиланнями на PDF-файли публікацій

Висновки

У запропонованій моделі ПрО як онтологічні зв’язки застосовуються зв’язки між областями інтересів окремих вчених. Фактично розглядається компактифікація біграфа «вчений – наукові поняття, які його цікавлять».

Запропоновано та реалізовано підхід до формування моделі ПрО, основу якого складають деякі маркери понять (теги), заздалегідь задані вченими (або, в окремих випадках, приписувані вченим) — учасниками проекту Google Scholar Citations.

Слід відзначити принципову відмінність запропонованої моделі автоматичного формування моделі ПрО від існуючих, що базуються на аналізі текстових корпусів (наприклад, [2]), або безпосередньої участі експертів при виборі конкретних вузлів і зв'язків [1]. Тут експерт-користувач вкладає лише крупинки знань у вигляді невеликих за обсягом словників тегів і шаблонів. Надалі автоматично використовуються знання, закладені самими авторами публікацій, теги відмічені ними як головні. Тобто експертне середовище в цьому випадку істотно розширюється.

Реалізовано алгоритм, відповідно до якого на підставі побудованої мережі формується бібліографічний список найбільш цитованих робіт у даній ПрО, що представлені у базі даних сервісу Google Scholar Citations.

Подібний підхід також може бути застосовано до бібліографічних баз даних, в яких у явному вигляді виділені автори і як теги — ключові слова.

Модель застосована для напрямку досліджень «екстрагування понять», але її можна використовувати і для інших наукових областей. Зокрема, вже побудовані подібні мережі для напрямків штучного інтелекту, багатоагентних систем і складних мереж (Complex Networks).

1. *Онтологии* и тезаурусы / Добров Б.В., Соловьев В.Д., Лукашевич Н.В., Иванов В.В. // Модели, инструменты, приложения. Бином, 2009. — 173 с.

2. *Ландэ Д.В*. Подход к созданию терминологических онтологий / Д.В. Ландэ, А.А. Снарский // Онтология проектирования. — 2014. — № 2(12). — С. 83–91.

3. *Чанышев О.Г*. Автоматическое построение терминологической базы знаний / О.Г. Чанышев // Труды 10-й Всероссийской научн. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008. — Дубна (Россия), 2008. — С. 85–92.

4. *Zeinalipour-Yazti D*. Information Retrieval in Peer-to-Peer Networks / D. Zeinalipour-Yazti, V. Kalogeraki, D. Gunopulos // IEEE CiSE Magazine, Special Issue on Web Engineering. — 2004. — P. 1–13.

5. *Kalogeraki V*. A Local Search Mechanism for Peer-to-Peer Networks / V. Kalogeraki, D. Gunopulos, D. Zeinalipour-Yazti // Proc. of CIKM'02, McLean VA, USA. — 2002.

6. *Yang B*. Efficient Search in Peer-to-Peer Networks / B. Yang, H. Garcia-Molina // Proc. of ICDCS'02, Vienna, Austria. — 2002.

7. *Erdős P*. On The Evolution of Random Graphs / P. Erdős, A. Rényi // Magyar Tud. Akad. Mat. Kutató Int. Közl. — 1960. — **5**. — P. 17–61.

8. *Albert R*. Statistical mechanics of complex networks / Réka Albert, Albert-László Barabási // Reviews of Modern Physics'74. — 2002. — P. 47–97.

9. *Ландэ Д.В*. Моделирование контентных сетей / Д.В. Дандэ // Проблеми інформатизації та управління: зб. наук. праць. — К.: НАУ, 2012. — Вип. 1(37). — С. 78–84.

Надійшла до редакції 15.09.2015