УДК 681.3

Олександр О. Марченко, к.ф.-м.н., доц.

Застосування алгоритму ієрархічної кластеризації слів у неевклідовому семантичному лексичному просторі для автоматизації побудови ієрархічних структур онтології

Стаття присвячена дослідженню моделі неевклідового семантичного лексичного простору та деталізації алгоритму ієрархічної кластеризації слів в даному просторі з ціллю побудови ієрархічних лінгвістичних структур онтологічного типу.

Ключові слова: семантична відстань, онтології, штучний інтелект.

Київський національний університет імені Тараса Шевченка, 03680, м. Київ, пр-т. Глушкова 4д, e-mail: rozenkrans@yandex.ua

Oleksandr O. Marchenko, Ph.D. (Physics & Mathematics), Associate Professor

Applying the algorithm of hierarchical clustering of words in lexical semantic non-Euclidean space to automate the construction of hierarchical structures of ontologies

The article describes researching of the model of non-Euclidean semantic lexical space and details the algorithm of hierarchical clustering of words in the space for the purpose of constructing hierarchical linguistic structures of ontological type.

Key words: semantic distance, ontologies, artificial intelligence.

Taras Shevchenko National University of Kyiv, 03680, Kyiv, Glushkova st., 4d, e-mail:*rozenkrans@yandex.ua*

Статтю представив чл.-кор. НАН України, д.ф.-м.н., проф. Анісімов А.В.

Визначення. Семантичним лексичним простором назвемо множину L точок – слів природної мови, з мірою семантичної близькості μ:L2 $\rightarrow$R, яка двом словам ставить у відповідність дійсне число – значення семантичної близькості між ними.

Найважливішим питанням для визначення семантичного лексичного простору є вибір міри семантичної близькості між словами. Традиційно для цього використовуються два підходи: міри, які обчислюються на таксономічних ієрархічних мережах онтологічних систем типу WordNet [1] та міри, що обчислюються за допомогою лексичного перетину тлумачних та енциклопедійних статей, що представляють собою визначення або опис слів-понять *x* та *y* (по типу міри Леска [2]), а також шляхом частотної та спеціальної обробки текстових корпусів. Типовими представниками першого підходу є такі міри, як міра Рєзника [3], міра Ву-Палмера [4] та міра Лікока-Ходорова [5]. Другий підхід широко представлений в роботах Мілн и Віттен [6] (метод Wikipedia Link-based Measure - WLM), у Габриловича и Марковича [7] (метод Explicit Semantic Analysis - ESA) , у Ех, Рамадж, Маннинг та ін. [8] (метод WikiWalk) , а також у А.В. Анісімова и др. [9] (метод Estimated Weighted Overlap − EWO). Сучасні дослідження показали, що латентний семантичний аналіз LSA великих текстових корпусів різноманітного тематичного змісту дає можливість побудови ефективних мір семантичної близькості між словами-термінами, як у [10].

Згідно канонічного визначення метричним простором є пара *(X, d)*, яка

© О.О. Марченко, 2013

складається з деякої множини елементів X і відстані $d : X \times X \rightarrow R$, а саме однозначної, невід'ємної, дійсної функції $d(x,y)$, визначеної для $\forall\ x,y \in X$, яка задовольняє наступні 3 умови:

1. $d(x,y) = 0 \Leftrightarrow x=y$ (аксіома тотожності)
2. $d(x,y) = d(y,x)$ (аксіома симетрії)
3. $d(x, z) \leq d(x,y)+d(y,z)$ (нерівність трикутника)

В лінгвістичному просторі з вищезгаданими мірами близькості між словами можна впевнено гарантувати виконання перших двох вимог класичного визначення метричного простору. Що стосується нерівності трикутника, то в даному випадку очевидно існують випадки порушення третьої умови по менший мірі з двох причин: по-перше, існують загальновживані слова, які для більшості наведених мір є однаково близько розташованими майже до всіх слів; по-друге, існує велика кількість багатозначних слів (наприклад слово «коса» має три значення – 1. зачіска; 2. інструмент; 3. довга смужка-півострів), і через такі слова опосередковано можна отримати набагато ближчу відстань між деякими семантично далекими одне від одного словами x та y, аніж пряма відстань між ними. Стосовно загальновживаних слів, то їх можна виключити з розгляду даної моделі, виявивши їх множину шляхом застосування частотного аналізу до великих текстових корпусів різної тематичної спрямованості із обчисленням оцінок значень TF-IDF[11] для слів в корпусах. Слова, які мають різні значення, є більш складною проблемою в плані усунення проблеми порушення нерівності трикутника. Пропонується при формуванні лексичного семантичного простору перейти від слів до їх семантичних значень. Множина L у визначенні може бути представлена як множина семантичних значень слів. Таким чином, слово ***коса*** має бути розділено на три його семантичних значення – ***коса1***, ***коса2*** та ***коса3***. Доцільно дослідити можливість автоматичного розщеплення слів на їх різні семантичні значення та переходу в простір семантичних значень слів. З'ясуємо, чи можна застосувати для цього алгоритми кластеризації у неевклідових просторах.

Розглянемо задачу кластеризації на неевклідовому семантичному просторі слів, як він введений на початку статті. Вона формулюється стандартним чином – необхідно розбити множину точок-слів на кластери згідно міри семантичної близькості. Тобто всередині кластеру слова мають бути ближчими один до одного ніж до слів з інших кластерів. Розглянемо ієрархічний підхід до кластеризації точок-слів у просторі.

Алгоритм ієрархічної кластеризації[12] має наступний вигляд:

WHILE ***it is not time to stop***
DO
pick the best two clusters to merge;
combine those two clusters into one cluster;
END

Для конкретизації алгоритму треба з'ясувати наступні деталі:

1. Як представляються кластери (через центроїди, кластероїди чи інше).
2. Критерій підбору двох кластерів, найкращих для злиття на поточний момент часу.
3. Критерій зупинки алгоритму.

Так як в даному випадку йдеться про неевклідовий простір, то представлення кластерів центроїдами є неможливим, тобто в задачі потрібно використовувати кластероїди для представлення семантичного кластеру – множини точок-слів, близьких за смислом. Не маючи можливості об'єднувати точки в неевклідових просторах, єдиним способом стає підбір слова-точки з кластеру для його представлення. В ідеалі, ця точка лежить максимально близько до усіх інших точок кластеру, тому в деякому сенсі вона лежить у «центрі» кластера. Називатимемо таку точку «кластероїдом». Вибирати кластероїд можна декількома шляхами, кожен з яких призначений мінімізувати

відстань між кластероїдом та іншими точками в кластері. Зазвичай кластероїдом вибирають точку, що мінімізує:

a. Суму відстаней до усіх інших точок у кластері.

b. Максимальну відстань до іншої точки у кластері.

c. Суму квадратів відстаней до інших точок у кластері.

Стосовно критерію підбору двох кластерів, кращих для злиття, то можна об'єднати два кластера з найближчими кластероїдами. Також в якості критерію можна використовувати найменшу середню мінімальну відстань між усіма парами точок з цих кластерів.

Інший критерій пов'язаний із вимірюванням щільності кластера, яка основана на визначенні *радіуса* або *діаметра* кластера. Обидва варіанти мають місце у неевклідовому просторі. Діаметр − це максимальна відстань між двома будь-якими точками у кластері. Радіус може буди знайдений при використанні кластероїду. Це відстань від кластероїду до найвіддаленішої з точок кластера.

В якості критерію вибору двох кластерів до об'єднання можна також зазначити об'єднання кластерів за мінімальним радіусом результуючого кластера. Інший варіант – за мінімальною середньою відстанню між кластероїдом та точками результуючого кластера. Ще один варіант – використовувати суми квадратів відстаней між точками результуючого кластера та його кластероїдом.

Також можна об'єднувати кластери за критерієм мінімального діаметру результуючого кластера. Можливі варіації та аналоги цього правила.

Існує декілька можливих критеріїв зупинки процесу кластеризації. Потрібно зупинити процес, коли найкраще об'єднання існуючих кластерів не є адекватним. Можна скористатися критерієм, що будь-який кластер не повинен мати середню відстань від кластероїду до його точок більшу за якийсь визначений ліміт. Цей підхід має місце, якщо є гарантія , що кластер не розтягнутий на велику відстань.

Можна продовжити кластеризацію, доки не утвориться один кластер. Таким чином генерується дерево, яке відображає послідовність, згідно якої точки простору об'єднувались. В даній задачі кластеризації множини слів в семантичному лексичному просторі можна представити утворене єдине дерево кластеризації як ієрархічну мережу деякої онтології.

Розглянемо детальніше критерії зупинки об'єднання. Існує критерій «зупинитись, коли залишилось k кластерів» у випадку, коли заздалегідь задане число k. Інші критерії:

1. Зупинитись, коли діаметр результуючого кластера перевищує порогове значення. Це також можна робити для радіуса.

2. Зупинитись, коли щільність результуючих кластерів буде менше якогось зазначеного порогу. Щільність можна задати різним чином. Наприклад, щільність обчислюється як кількість точок, поділених на потужність діаметру або радіусу кластера.

3. Зупинитись, коли є підстави вважати результуючий кластер наступного об'єднання поганим. Наприклад, можна відстежувати середній радіус чи діаметр поточних кластерів. Якщо об'єднати два кластери, які не повинні об'єднуватись, тоді середній радіус/діаметр різко зросте відносно попередніх темпів росту радіусу/діаметру. При цьому відбудеться різке падіння щільності утвореного кластера відносно попередніх кластерів.

Як вже було зазначено, можна продовжити кластеризацію, доки не утвориться один кластер для утворення ієрархічної мережі об'єднання точок-слів, яку можна проінтерпретувати як ієрархію деякої онтології.

При цьому в процесі переобчислення кластероїдів та відстаней від них до всіх інших елементів цього кластеру можна відносно просто виявити такі слова, які одночасно однаково або майже однаково наближені до кількох кластероїдів

одночасно. Пряма відстань між кластероїдами є набагато довша, аніж сума відстаней між цими кластероїдами та таким проміжним словом. Таке порушення нерівності трикутника вказує на багатозначність проміжного слова. Воно має бути розділено на ту кількість значень, яка відповідає числу кластероїдів, квазірівно-наближених до цього слова. Тому до алгоритму ієрархічної кластеризації на кожній ітерації переобчислення кластероїдів додається спеціальна процедура перевірки та виявлення багатозначних слів з подальшим їх розщепленням та зарахуванням утворених різних значень до відповідних квазі-рівноблизьких кластерів. Потребує уточнення ще і критерій вибору кластероїду: кластероїдом можна обирати лише однозначне слово – тобто для нього має виконуватися нерівність трикутника в середині кластера: не може бути, щоб відстань між будь-якими двома словами з даного кластеру була більша, аніж сума відстаней між ними та кластероїдом.

Ієрархічна кластеризація може утворити каркас онтологічної мережі лінгвістичної бази знань. Очевидно, що для цього доцільно використовувати міри семантичної близькості, побудовані за допомогою методів дистрибутивної семантики – наприклад, шляхом застосування методу латентного семантичного аналізу LSA на великому текстовому корпусі з тематичним покриттям значної кількості напрямків.

Таким чином, дослідження показали значні перспективи розробки механізмів автоматичної побудови ієрархічних онтологічних структур лінгвістичних баз знань на основі алгоритмів ієрархічної кластеризації, застосованих до семантичного лексичного простору слів з мірами семантичної близькості, створеними методом латентного семантичного аналізу LSA на великих текстових корпусах.

Список використаних джерел

1. *G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller* WordNet: An online lexical database. / G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller //Int. J. Lexicograph. – 1990.-3, 4, pp. 235–244

2. *Lesk,M.* Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In SIGDOC '86: Proceedings of the 5th annual, 1986.

3. *Resnik, P.* Using information content to evaluate semantic similarity in a taxonomy. In International Joint Conference for Artificial Intelligence (IJCAI-95), 1995, pages 448-453.

4. *Wu, Z. and Palmer, M.* Verb semantics and lexical selection. In 32nd. Annual Meeting of the Association for Computational Linguistics, 1994, pages 133-138, New Mexico State University, Las Cruces, New Mexico.

5. *Leacock, C., Chodorow, M., and Miller, G. A.*. Using corpus statistics and wordnet relations for sense identification. Computational Linguistics, 1998, 24(1):147-165.

6. *Milne D., Witten I.H.* An effective, low-cost measure of semantic relatedness obtained from wikipedia links. Association for the Advancment of Artifical Intelligence, 2008

7. *Gabrilovich E., Markovitch S.* Computing semantic relatedness using Wikipedia-based explicit semantic analysis. Department of Computer Science Techion - Israel Institute of Technology. 2006.

8. *Yeh E., Ramage D., Manning C. D., Agirre E., Soroa A.* WikiWalk: Random walks on Wikipedia for Semantic Relatedness. ACL-IJCNLP TextGraphs-4 Workshop 2009.

9. *Anisimov, A.; Marchenko, O.; Kysenko, V.* A method for the computation of the semantic similarity and relatedness between natural language words:- Cybernetics and Systems Analysis, Volume 47, Number 4, July 2011 , pp. 515-522(8)

10. *Marchenko, O.* A method for automatic linking of Wikipedia entries as new nodes of Ukrainian ontological knowledge base UkrWordNet:- Bulletin of Taras Shevchenko National University of Kyiv: Series Physics & Mathematics:-№1:-2013

11. *Salton, G. and Buckley, C.* 1988 Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5): 513—523

12. *Anand Rajaraman, Jeffrey David Ullman* Mining of Massive Datasets:- Cambridge University Press:-2012:-p.328

Надійшла до редколегії 18.04.2013