УДК 004.912:519.766.4

А.А. Дегтярьов, Т.А. Зайцева
Дніпропетровський національний університет ім. Олеся Гончара

СИМПЛІСТИЧНИЙ МЕТОД ТЕМАТИЧНОЇ ІДЕНТИФІКАЦІЇ НАТУРАЛЬНОМОВНОГО ТЕКСТУ НА ОСНОВІ РОЗПОДІЛУ КЛЮЧОВИХ ТЕРМІВ У ТЕКСТІ

Пропонується спосіб оцінки ступеня належності тексту до певної, наперед заданої тематики. Тематичний індекс розраховується як дійсне число та може бути використаний для порівняння з деяким пороговим значенням при визначенні тематичної сутності тексту. Підхід базується лише на припущенні, що задана множина характеристичних слів для оцінюваного тематичого напрямку. Запропонований метод має просту реалізацію, яка підходить як для розв'язку задач, де висока точність оцінки не є пріоритетною вимогою, так і для розв'язку більш складних задач, де даний метод може бути використаний у контексті попередньої оцінки тексту у багатоетапному процесі тематичної ідентифікації.

Ключові слова: *обробка натуральних мов, тематична ідентифікація, розпізнавання тематичного напрямку, тематичний індекс, оцінка тематичної належності тексту, автоматична каталогізація текстів.*

Предлагается метод оценки степени принадлежности текста на натуральном языке некоторой, наперед заданной тематике. Тематический индекс рассчитывается как действительное число и может быть использован для сравнения с некоторым пороговым значением при определении тематической сущности текста. Подход основывается только на предположении о том, что задано множество характеристических слов для оцениваемой тематики. Метод подразумевает простую реализацию, подходящую как для решения задач, где высокая точность оценки не является доминирующим требованием, так и для решения более сложных задач, где предлагаемый метод может быть использован в контексте предварительной оценки текста в многоэтапном процессе тематической идентификации.

Ключевые слова: *обработка естественных языков, тематическая идентификация, распознавание тематического направления, тематический индекс, оценка тематической принадлежности текста, автоматическая каталогизация текстов.*

A simplistic method of topic identification is proposed. The topic index is calculated as a real number, and can be used for comparing with a threshold when validating the topic identity of the content. The approach is based on the assumption of having a set of characteristic words for the estimated topic and yields a simple implementation which is suitable both for the tasks where the high precision is not required and for solving more complex problems, where it can be used in the context of preliminary topic evaluation in a multi-level topic identification process.

Key words: *natural language processing, topic identification, topic recognition, topic index, estimating the relevancy of a text in the context of a given topic, automatic text categorization.*

Постановка проблеми. Проблема визначення тематичної спрямованості отриманого на вхід довільного тексту постає у задачах автоматичної каталогізації ресурсів. Зокрема, ця задача може виникати у різноманітних інтернет-додатках, де вона тісно пов'язана із задачею валідації контента на відповідність певним вимогам, що висуваються до ресурсу. Наприклад, задача тематичної ідентифікації стає актуальною при динамічному наповненні індексної бази певної галузевої пошукової системи, коли необхідно визначити, чи необхідно заданий документ включати до індексу або відкинути його як такий, що не відповідає профілю системи. Також, задача тематичної валідації може застосовуватись при перевірці коректності нового повідомлення на форумі і визначення відповідності змісту повідомлення загальній тематиці форуму, на основі чого системою буде прийняте рішення про публікацію або відкладення його для подальшої перевірки.

Аналіз останніх досліджень і публікацій. Існує ряд підходів до тематичної ідентифікації тексту [4; 5], в межах кожного з яких робиться акцент на певному аспекті тематичної складової текстового потоку. У той час, як практично кожен з них базується на концепції наборів ключових слів, що характеризують певний тематичний напрямок, методи подання цих наборів і встановлення зв’язку між ними та вхідним текстом у контексті його аналізу істотно відрізняються [6; 7]. Варто зазначити, що у сфері тематичної ідентифікації тексту слід розділяти задачі, які сфокусовані на побудові наборів ключових слів для заданих тематичних напрямків, та задачі, які сфокусовані саме на визначенні ступеня належності вхідного тексту певному тематичному напрямку, який задається вже визначеним набором ключових слів.

Методи тематичної класифікації тексту можна розділити на дві групи: лінгвістичні [5] та статистичні [1; 4]. Крім того, існують розроблені методи, що спираються на певні лінгвостатистичні

особливості тексту, зумовлені його авторством [3], або специфікою предметної області [2]. У кожній з цих груп існують досить точні методи, але їх реалізація з метою практичного застосування може вимагати певного об'єму ресурсів, який виявляється необґрунтовано великим при застосуванні для вирішення відносно простих задач. У цьому контексті виникає необхідність у розробці досить простого в реалізації методу, який не потребував би великих витрат як машинних (підготовка та зберігання великого об'єму лінгвістичних даних, час виконання, апаратні ресурси), так і людських (людиногодини розробника програмного забезпечення) ресурсів.

Постановка задачі. Нехай $C = \{w_1, w_2, \ldots, w_n\}$ – вхідний контент, заданий у вигляді скінченої послідовності слів w_i, $i = \overline{1, n}$ певної мови L. Нехай також $D(T) = \{t_1, t_2, \ldots, t_k\}$ – заданий тематичний словник теми T. Тематичним словником теми T будемо називати скінчену множину таких слів t_i, $i = \overline{1, k}$ мови L, які є характеристичними для теми T, тобто поява кожного з них у тексті підвищує імовірність того, що даний текст належить до тематичного напрямку T. Формально це може означати, що питома вага тематичного слова в тексті, про який заздалегідь відомо, що він належить до тематичного напрямку T, є у середньому вищою за питому вагу інших слів, що з'являються у тексті. Тематичність слова може також визначатись тим, що його частота у тематичних текстах вища за частоту, з якою це слово зазвичай з'являється в мові L. Наприклад, у контексті натуральних мов, значне відхилення частоти слова у певному наборі тематичних текстів від значення, що визначено для нього в мовному корпусі, може бути підставою для внесення його до тематичного словника цього тематичного напрямку. Слід сказати, що задача формування наборів характеристичних слів є окремою за своїм характером, і в задачі визначення належності вхідного текстового потоку до певного тематичного напрямку ми будемо вважати тематичний словник для цього напрямку вже сформованим та заданим заздалегідь.

Сформулюємо задачу визначення індексу належності тексту до заданого тематичного напрямку наступним чином: необхідно визначити таку функцію $I: (C, D(T)) \to [0; 1]$, яка ставить у відповідність вхідному текстовому потоку певне чисельне значення індексу належності до тематики T за заданим тематичним словником $D(T)$.

Основний матеріал. Концептуальний підхід до розрахунку тематичного індексу полягає у поданні вихідної величини як добутку складових, які також належать проміжку [0; 1], та відображають ступень вираженості тієї чи іншої характеристики щодо властивостей розташування тематичних слів за контентом. Таке розбиття на компоненти дозволяє звести обчислення досить абстрактної величини до обчислення цілком конкретних характеристик текстового потоку, на основі чого можна також робити висновки щодо певних особливостей контента, що розглядається, та проводити додатковий аналіз з метою виявлення інших закономірностей, що можуть представляти інтерес для дослідження в цілому.

Характеристики, що обираються для обчислення сумарного тематичного індексу, повинні бути в певному сенсі незалежними одна від одної, та виявляти різні аспекти вхідного текстового потоку з боку особливостей появи тематичних слів у ньому. Наприклад, очевидно, що обрання лише однієї характеристики, що обчислюється як питома вага тематичних слів у контенті, не може давати вичерпної картини щодо тематики усього контента, бо тематичні слова можуть бути сконцентровані лише у кількох реченнях, де коротко йдеться про предметну область, для якої ці слова є характерними, у той час як весь він у цілому належить до іншої предметної області. У такому випадку, очевидно, треба також враховувати і рівномірність появи слів по всьому текстовому потоку, щоб виключити сильний вплив випадкової появи кількох з них лише в одному фрагменті.

У цьому контексті, в якості характеристик для розрахунку сумарного тематичного індексу текстового потоку, було обрано наступні показники:

- Кількісна оцінка ваги тематичних слів по відношенню до всього контента.
- Оцінка рівномірності розташування тематичних слів у конвенті.
- Оцінка ступеня унікальності набору тематичних слів, що зустрічаються у контенті (оцінка різноманітності представлених на контенті тематичних слів).

Не зважаючи на однобічність, яка властива кількісній оцінці питомої ваги тематичних слів, вона повинна впливати на сумарне значення індексу, бо більш висока частота появи тематичних слів у тексті непрямо підвищує імовірність належності цього тексту до відповідного тематичного напрямку. При обчисленні цієї величини, важливо зазначити, що як правило, при прямому підрахунку, питома вага тематичних слів є досить малою величиною, з погляду на те, що

фактично кожна натуральна мова містить велику кількість службових слів та конструкцій, які не несуть певної тематичної завантаженості, але своєю появою у тексті значно примножують сумарну кількість слів у ньому. Тому, задля поліпшення обчислень, та відходу від замалих величин, має сенс ввести певний істотний поріг питомої ваги тематичних слів, та розраховувати значення першої характеристики відносно цього порогу. Значення порогу повинно бути водночас досить великим, щоб тексти, що не є тематичними, давали значення цієї оцінки близьким до нуля, та досить малим, щоб більшість тематичних текстів отримували значення цієї оцінки близьким до одиниці. Виходячи з цього, першу оцінку пропонується обчислювати наступним чином

$$q(D(T);C)=\begin{cases}\dfrac{N'}{k_q N}, & \dfrac{N'}{N}<k_q\\ 1, & \dfrac{N'}{N}\geq k_q\end{cases}, \qquad (1)$$

де N – загальна кількість слів у контенті C, N' – кількість тематичних слів у контенті C (або кількість елементів у послідовності $Q=\{v:(v\in C)\wedge(v\in D)\}$), k_q – параметр, який визначає істотний поріг питомої ваги тематичних слів. Значення порогу, у загальному випадку, залежить від мови, та визначається особливостями тематичних текстів, поданих в якості бази для автоматичного навчання системи. Наприклад, значення порогу може бути обчислено як середнє значення питомої ваги тематичних слів у цих текстах, про які заздалегідь відомо, що вони належать певній тематиці. Емпірічно було встановлено, що у більшості випадків оптимальним значенням є $k_q=0.15$.

Значення другої обраної характеристики, а саме оцінки рівномірності розташування тематичних слів у контенті, пропонується обчислювати базуючись на підході, описаному у [7] для підрахунку рівномірності розташування тегів частин мови у реченні. Після адаптації у даному контексті, підсумкова формула виглядає наступним чином:

$$d(D(T);C)=1-\mu W(D(T);C), \qquad (2)$$

де $W(D(T);C)$ – величина, яку назвемо дисперсією тематичних слів із словника $D(T)$ у контенті C, та яка обчислюється за наступною формулою

$$W(D(T);C) = \frac{1}{N'+1}\sum_{i=0}^{N'}\left(d_i - \frac{N-N'}{N'+1}\right)^2. \quad (3)$$

У загальному випадку, ця величина потребує нормалізації до інтервалу [0; 1] задля використання як множника при підрахунку тематичного індексу. Коефіцієнт нормалізації μ виконує цю функцію, та обчислюється наступним чином

$$\mu = \frac{1}{N'}\left(\frac{N'+1}{N-N'}\right)^2. \quad (4)$$

У формулах (3) та (4) величини N та N' визначені так само, як в (1), а d_i - відстань у словах між двома послідовними (i-ою та $(i+1)$-ою) появами тематичних слів у контенті, причому d_1 – це кількість слів від початку тексту до першої появи тематичного слова, а $d_{N'+1}$ – кількість слів від появи останнього тематичного слова в контенті до його кінця. Тривіальний випадок $N'=0$ не розглядається, бо в такому випадку сумарний тематичний індекс можна вважати нульовим, без обчислення жодної із характеристик. Формула (2) випливає з наступного твердження: чим менша сума квадратів різниць відстаней між послідовними появами тематичних слів у контенті та величиною, що виражає відстань між тематичними словами при ідеальній рівномірності, тим рівномірнішим є розташування слів за контентом, і тим більшим повинне бути значення другої характеристики. Коефіцієнт нормалізації обчислюється виходячи з найгіршого з боку рівномірності випадку: якщо всі тематичні слова розташовані послідовно на початку або в кінці контента, то вектор відстаней має вигляд $\left(\underbrace{0,0,0,...,0}_{N'},(N-N')\right)$, а міра рівномірності вважається рівною нулю, тобто коефіцієнт μ фактично визначається з рівняння $d(D(T);C)=0$.

Третьою характеристикою, яку пропонується обчислювати та включати до складу компонент сумарного тематичного індексу, є міра унікальності тематичних слів у наборі, що з'являється у контенті. Ця характеристика виражає те, наскільки різноманітним є набір тематичних слів на контенті. Урахування цієї міри дозволяє виключити сильний вплив випадкової, але частої появи лише кількох тематичних слів у нетематичному контенті, що може виникати, наприклад, при невирішених омонімічних колізіях у тематичному

словнику. Чисельне значення цієї характеристики пропонується обчислювати наступним чином

$$u(D(T);C)=\begin{cases}\dfrac{N''}{k_u N'}, & \dfrac{N''}{N'}<k_u\\ 1, & \dfrac{N''}{N'}\geq k_u\end{cases}, \qquad (5)$$

де N та N' визначені так само, як в (1), N'' – кількість унікальних тематичних слів у контенті C (тобто кількість всіх слів контента, які належать тематичному словнику $D(T)$, без урахування дублікатів), а k_u – істотний поріг унікальності, який введено з таких міркувань, як і істотний поріг питомої ваги у (1). Значення порогу може коливатись та є менш чутливим з погляду впливу на об'єктивне формування сумарного тематичного індексу. Значення порогу залежить від предметної області, характеру тематичного словника та його потужності, а також мови контента, що оцінюється. За результатами проведених експериментів для різних крайових випадків, було встановлено, що адекватне значення цього параметра лежить у проміжку [0.3; 0.6]. Точне значення порогу повинно обиратись залежно від обсягу тематичного словника, а також виходячи із специфіки предметної області та особливостей уживання ключових термів у типовому контенті, що належить даній тематиці.

Після обчислення кожної з наведених характеристик, сумарний тематичний індекс контенту обчислюється за наступною формулою

$$I(D(T);C)=q(D(T);C)*d(D(T);C)*u(D(T);C). \qquad (6)$$

Слід зазначити, що у певних випадках, з метою оптимізації часу та пам'яті, що витрачаються при подальших обчисленнях з використанням значення сумарного тематичного індексу, має сенс зводити обчислене значення до цілочисельного інтервалу [0; 255] та використовувати його у подальшому.

Проілюструємо розрахунок тематичного індексу за допомогою запропонованого методу на прикладі першого абзацу частини 1 роботи [5]. В якості тематичного напрямку, належність до якого підлягає аналізу, будемо розглядати обробку натуральних мов (англ. Natural Language Processing), а в якості тематичного словника оберемо множину специфічних для даної тематики слів. Базовий набір тематичних слів може бути складений на основі енциклопедичної статті, яка дає широке означення обробки натуральних мов як області штучного інтелекту (прикладом може бути стаття з вікіпедії). Після

первісного аналізу вхідного тексту можна скласти наступний набір слів, які з високою імовірністю будуть присутніми у кожному тематичному словнику за даною тематикою:

$$D(T) = \{"linguistics","syntactic","word", "model","grammar","text","sentence", "graph","relation","subject","meaning","verb"\} \quad (7)$$

Для проведення аналізу корисно побудувати карту розташування слів за контентом. Зазначимо, що для підвищення точності статистичного аналізу, до загальної карти має сенс включати лише ті слова, які можуть істотно впливати на загальний зміст тексту, і, відповідно, виключати всі службові елементи. До службових елементів, у першу чергу, слід віднести всі цифрові слова (роки, номери тощо), та деякі службові частини мови (сполучники, займенники, прийменники, частки та артиклі). Така фільтрація є простою у реалізації, так як списки слів, що належать до другорядних частин мови, досить малі, повністю відомі, та практично позбавлені необхідності розв'язання омонімічних колізій. Після фільтрації, можна побудувати наступну карту розташування слів (тематичні слова виділені спеціальним чином):

Карта розподілу тематичних слів по контенту

Syntactic	dependency	representations	long	history	descriptive	theoretical	**linguistics**
Many	formal	**models**	advanced	most	notably	**Word**	**Grammar**
Hudson	**Meaning**	**Text**	Theory	Melchuk	Functional	Generative	Description
Sgall	Hajicov	Panevov	Constraint	dependency	**Grammar**	Maruyama	Common
Common	theories	notion	directed	**syntactic**	dependencies	**words**	**sentence**
Example	given	Figure	**sentence**	hearing	scheduled	issue	today
Extracted	Penn	Treebank	Marcus	Santorini	Marcinkiewicz	dependency	**graph**
Sentence	represents	**word**	**syntactic**	modifiers	labeled	directed	arcs
Arc	label	comes	finite	set	representing	possible	**syntactic**
Roles	Returning	Example	figure	see	multiple	instances	labeled
dependency	**relations**	finite	**verb**	hearing	labeled	SBJ	indicating
Hearing	head	**syntactic**	**subject**	finite	**verb**	artificial	**word**
Inserted	beginning	**sentence**	always	serve	single	root	**graph**
Primarily	means	simplify	computation				

За даною картою розташування загальна кількість слів, що підлягає аналізу, дорівнює $N = 108$, кількість тематичних слів становить $N' = 25$, а кількість унікальних слів у тематичному наборі становить $N'' = 12$. Проводячи розрахунок розглянутих характеристик за формулами (1) , (2), (5), отримуємо наступні результати:

$q(D(T); C) = 1$, так як $N / N' = 0.2315 > k_q = 0.15$;

$$d(D(T);C) = 1 - \frac{1}{25} * \left(\frac{26}{108-25} \right)^2 * \frac{1}{26} * \sum_{i=1}^{26} \left(d_i - \frac{108-25}{26} \right)^2 = 0.9502$$

при векторі відстаней визначеному з карти як $(0, 6, 2, 3, 0, 1, 0, 10, 6, 1, 0, 3, 11, 0, 1, 0, 11, 9, 1, 6, 0, 1, 1, 2, 4, 4)$;

$u(D(T);C)$, так як $N / N' = 12/25 = 0.48 > k_u = 0.45$.

Сумарна оцінка ступеня належності поданого тексту до розглянутої тематики дорівнює:

$$I(D(T);C) = q(D(T);C) * d(D(T);C) * u(D(T);C) = 1 * 0.9502 * 1 = 0.9502.$$

Обчислене значення слід розглядати як досить високе та ідентифікувати текст як такий, що належить тематиці, представленій розглянутим тематичним словником.

На практиці важливим вдосконаленням такого підходу є врахування можливості присутності на контенті слів, що належать до іншої тематики або є небажаними. У такому разі тематичний індекс контенту являє собою алгебраїчну суму значень, обчислених за вищенаведеним принципом, взятих із відповідним знаком, при цьому множина тем розділяється відповідно на дві підмножини. Також важливим вдосконаленням є розділення вхідного потоку на основний текст та підзаголовки, що можливо практично для кожного популярного формату. При такому уточненні сумарний тематичний індекс контенту розраховується як середнє зважене окремо розрахованих індексів за основним текстом та підзаголовками, причому підзаголовки можуть отримувати більшу вагу.

Після розрахунку значення тематичного індексу, необхідно прийняти рішення про валідацію контента як такого, що підходить або не підходить вимогам, висунутим у конкретній задачі. Для цього заздалегідь доцільно задати значення порогу тематичного індексу, при подоланні якого відповідний ресурс визнається валідним.

Висновки. У результаті проведеного дослідження запропоновано метод оцінки належності вхідного натуральномовного тексту до певного тематичного напрямку, заданого набором нормалізованих ключових термів. У контексті запропонованого методу побудовано функцію розрахунку тематичного індексу (7), яка представляє собою добуток нормалізованих множників, що визначаються за формулами (1), (2), (5), та семантично представляють основні характеристики розподілу ключових термів у вхідному текстовому потоці (кількість, рівномірність та унікальність).

Запропонований підхід до розв'язку задачі тематичної валідації тексту є досить універсальним, і може бути використаний у силу своєї швидкісності практично у всіх проектах, де швидкість є пріоритетною вимогою. Серед таких задач варто відзначити побудову системи інформаційного пошуку, зокрема системи автоматичного супроводження процесу огляду літератури під час виконання роботи студентами, фахівцями, аспірантами та науковими співробітниками, в ході якого виконується автоматична каталогізація літератури та уточнення набору тематик у процесі пошуку певного питання.

Бібліографічні посилання

1. **Steyvers M., Griffiths T.** Probabilistic topic models. / Latent Semantic Analysis: A Road to Meaning – University of California, Irvine, 2007. – P. 1–15.

2. **Andrzejewski D., Zhu X., Craven M., Recht B.** A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation using First-Order Logic // IJCAI – 2011.

3. **Dai M. A., Storkey A. J.** The Grouped Author-Topic Model for Unsupervised Entity Resolution. // ICANN – 2011.

4. **Singhal A., Mitra M., Buckley C.** Learning routing queries in a query zone. // In Proc. of the SIGIR'97. – 1997. – P. 25–32.

5. **Hatzivassiloglou V., Gravano L., Maganti A**. An investigation of linguistic features and clustering algorithms for topical document clustering. // In Proc. of the SIGIR'2000. – 2000.

6. **Salton G., Allan J., Singhal A.** Automatic text decomposition and structuring. // Information Processing & Management. – 1996. – 32(2) – p. 127–138.

7. **Salton G., Singhal A., Mitra M., and Buckley C.** Automatic text decomposition and summarization. // Information Processing & Management. – 1997. – 33(2) p. 193–208.

8. **McDonald R., Nivre J.** Analyzing and Integrating Dependency Parsers. // Computational Linguistics. – vol. 37. – 2011.

9. **Дегтярьов А. А.** Контекстно-зважена тематична валідація тексту довільного характеру. /А. А. Дегтярьов // Системний аналіз та інформаційні технології: Матеріали 12-ї міжнародної науково-технічної конференції SAIT 2010, Київ, 25-29 травня, 2010 р. – К., 2010. – 544 с.

10. Дегтярёв А. А. Лингвостатистическая модель оценки репрезентативности текстового фрагмента с применением адаптационных механизмов определения границ предложения. / А. А. Дегтярёв, Т. А. Зайцева // Питання прикладної математики і математичного моделювання. – Дніпропетровськ, 2012. – С. 94–102.

Надійшла до редколегії 11.07.2012