УДК 004.774.6.: 001.82

Олександр Кузнєцов,
науковий співробітник відділу
довідково-консультаційної допомоги
Національної юридичної бібліотеки

Андрій Кузнецов,
провідний бібліотекар НБУВ

Контент-аналіз як засіб виявлення ступеня новизни публікацій

Наведено методику визначення тематичної близькості та ступеня новизни публікацій на основі контент-аналізу. Стаття одного автора порівнюється зі статтями іншого автора одного тематичного напряму за частотним використанням слів.

Ключові слова: *контент-аналіз, слово, текст, корпус, інформація, мова.*

Постановка проблеми. У наш час обсяг наукової інформації кожні п'ять років зростає вдвічі. Склалася ситуація, коли майже 95% всієї наукової інформації залишаються невикористаними. Причина насамперед полягає в обмеженій швидкості її опрацювання людським мозком (десь приблизно 50 біт у секунду). Зрозуміло, що для отримання корисної інформації потрібен різноманітний інструментарій [1].

Сьогодні структура й обсяги інформаційних потоків, у яких доводиться вишукувати крихти необхідної, готової до безпосереднього використання інформації, зумовлюють особливості самого процесу пошуку [2]. З появою дешевих комп'ютерів та доступного Інтернету актуальність якості інформаційного простору зростає. Розмістити інформацію можна без усіляких перешкод у необмеженій кількості. У різних наукових фахових виданнях України з'являються майже однакові статті або статті з однаковими ідеями, концепціями, що робить науковий простір біднішим. Ще одна з причин інформаційного колапсу — зростання обсягів неякісної інформації.

У системі наукових комунікацій періодичні видання, що становлять близько третини документального інформаційного потоку, виконують низку функцій, без яких неможливий розвиток науки. Стаття у науковому журналі має особливе значення: у ній не просто фіксується знання, вона стає необхідною формою закріплення і трансляції нового наукового результату та визначає "прозорість" авторського права [3].

Виявити новизну — корисну інформацію з текстів (статей) — можна за допомогою контент-аналізу. Контент-аналіз (якісно-кількісний аналіз) — формалізоване дослідження змісту тексту для визначення одиниць, які характеризують ту частину інформації, що стосується мети дослідження, а також систематизація та узагальнення кількісної інформації для виявлення причиново-наслідкових зв'язків. Очевидно, що для автоматизованих систем найзручнішим є застосування згаданого методу [4].

Аналіз останніх досліджень. Існують потужні комерційні програмні продукти для інформометричних досліджень, наприклад: IN-SPIRE™ Visual Document Analysis для опрацювання текстових даних науково-технічної літератури; VantagePoint для глибинного аналізу тексту.

Альтернатива платним продуктам — велика кількість вільно поширюваних програм: HistCite, Bibexcel, CiteSpace, Loet Leydesdorff тощо. HistCite — програмний продукт, призначений для формалізованого аналізу результатів пошуку в БД Web of Science, що містять ключові слова, авторів і процитованих авторів, журнали, країни й організації, від яких ті публікують свої статті. Bibexcel використовується для аналізу текстових даних, відформатованих для імпорту в Excel або іншу програму, що працює з табличними даними, для подальшого опрацювання. CiteSpace підтримує структурний і часовий аналіз різних мереж із наукових документів. Формат текстових файлів для аналізу такий, як у HistCite і Bibexcel. Loet Leydesdorff є набором програм для розбирання, перетворення й аналізу бібліометричних даних, отриманих з таких джерел, як Scopus, Web of Science і Google Scholar. Можна проводити аналіз на співавторство, мережі спільних проектів між країнами, організаціями та містами, аналіз на однакові ключові слова, співцитування, бібліографічний аналіз тощо. Ці програми не містять інструментів візуалізації, однак є можливість для створення реляційної бази даних для візуалізації в інших програмах. На жаль, у таких системах користувач-дослідник обмежений рамками можливостей відповідної системи та вимогами до вхідних текстових файлів (наявність структурованих бібліографічних даних) [5].

Мета статті — запропонувати методику виявлення новизни на основі контент-аналізу в неструктурованих текстових масивах (статтях).

Викладення основного матеріалу. Тематична близькість документів визначається тим, наскільки часто терми, тобто слова, використовувались у документах (статтях) [6].

Розглянемо для прикладу статтю:

Воєводська В., Войцехович А., Котіков Ю. Рекомендації щодо транслітерування літерами української абетки власних назв, поданих англійською, французькою, німецькою та італійською мовами // Офіційний веб-портал Державної служби інтелектуальної власності України / В. Воєводська, А. Войцехович, Ю. Котіков, Н. Куземська, В. Моргунюк, А. Новікова, Л. Пшенична, Л. Шрамко. — Режим доступу: http://sips.gov.ua/ua/transliteruvannja.html. — Назва з екрана.

Статті, з якими треба порівняти цю статтю, можна вважати корпусом. Корпус текстів — це вид корпусу даних, одиницями якого є тексти або їхні значні частини, що включають, наприклад, якісь повні фрагменти макроструктури текстів цієї проблемної галузі.

Корпус текстів характеризується чотирма основними параметрами: по-перше, він має бути великого обсягу; по-друге — структурованим або розміченим; по-третє, тексти, складові певного корпусу, — в електронному варіанті; по-четверте, в поняття "електронний корпус" входить, як правило, спеціальне програмне забезпечення для роботи з цим корпусом.

Цінність корпусу вбачається ось у чому:

— одного разу зроблений корпус може використовуватися багато разів;

— корпус показує мовні дані в їхньому реальному оточенні, що дає змогу досліджувати лексичну і граматичну структури мови, а також безперервні процеси змін, що відбуваються у ній упродовж певного відрізка часу;

— корпус характеризується збалансованим складом текстів, що уможливлює його використання для тестування пошукових машин, машинних морфологій, систем перекладу, а також у різних лінгвістичних дослідженнях;

— корпус має важливе значення для викладання мови, оскільки з його допомогою можна швидко та ефективно перевірити особливості вживання незнайомого слова або граматичної форми.

Робота з корпусами, тобто з масивами текстів в електронному форматі, давно вже стала одним з основних методів лінгвістичних досліджень. Статті, що відповідають визначенню корпусу:

1. Vakulenko M. O. Transliteration through a slavonic latin alphabet: saving information and expenses / M. O. Vakulenko // Вісник Київського лінгвістичного університету. Серія "Філологія". — 1999. — Т. 2, № 1. — С. 85—94.

2. Вакуленко М. О. Про те, яка латиниця нам потрібна, і трохи про "заочність" / М. Вакуленко, О. Вакуленко, О. Білодід, М. Корнілов // Слово Просвіти. — 1996. — Ч. 9/10.

3. Вакуленко М. О. Восточнославянская латиница в международном контексте / Maksim Olegovič Vakulenko // Slavia, Praha. — 1998. — R 67. — С. 333—339.

4. Вакуленко М. О. Наукові засади відтворювання запозичених та іншомовних слів: інваріантна транскрипція і транслітерація / Максим Вакуленко // Вісник Книжкової палати. — 1999. — № 10. — С. 6—9.

5. Вакуленко М. О. Наукові засади відтворювання запозичених та іншомовних слів: інваріантна транскрипція і транслітерація / Максим Вакуленко // Вісник Книжкової палати. — 1999. — № 11. — С. 15—18.

6. Вакуленко М. Про "складні" проблеми українського правопису (укр. латиниця, запозичені слова та ін.) / М. Вакуленко — К. : Курс, 1997. — 32 с.

Програмне забезпечення "Аналіз текстів" обробляє кожну статтю таким чином:

1. З тексту статті видаляються усі службові символи.

2. Видаляються "стоп-слова": англійські, російські та українські. Стоп-слова — це слова, які не несуть інформаційного навантаження щодо змісту статті, тобто на них можна не звертати увагу.

3. Нормалізуються ключові слова за словником синонімів: усі слова-синоніми замінюються одним словом також трьома мовами.

Використання трьох мов у масивах стоп-слів та синонімів зумовлено тим, що в наукових фахових виданнях України та СНД ці мови використовуються найчастіше і разом. Спосіб заміни деяких слів на відповідні синоніми здебільшого є ефективним для авторів, які переробляють чужі публікації, створюючи свої, проте насправді новизни у їхніх статтях немає. Далі будуються два частотних словники: статті та корпусу. Потім аналізуються два масиви, кожен з яких має таку структуру: слово та частота його використання у тексті. Визначаються масиви слів: всі слова статті; всі слова корпусу; слова, які є в обох масивах; унікальні слова статті; унікальні слова корпусу. Кількість слів, які збігаються і у статті, і у корпусі, — 126, що становить більше ніж чверть загальної кількості слів статті (365). Можна зробити висновок про один тематичний напрям та необхідність подальшого аналізу для виявлення ступеня новизни.

Розглянемо отриману інформацію, викладену в стовпцях (стаття, корпус) як вектори, та застосуємо до неї векторний аналіз. Найпридатнішим у цьому випадку є метод, що базується на прикладах. Якщо документи подібні — одного напряму, — то можемо застосувати косинусну міру, що дає можливість обчислювати відстань між документами та порівнювати тексти і вектори, які їм відповідають. Тобто від простору слів переходять до простору векторів, координатами яких є певні частотні показники слів — прості числа.

На практиці показник схожості (косинусна міра) для авторського тексту з науковою новизною має відповідати проміжку з величинами від 0,3 до 0,5. Для двох однакових текстів цей показник дорівнюватиме одиниці. Показник від 0,3 до 0,5 гарантує, що досліджувана стаття має певну наукову цінність і заслуговує на увагу. У нашому випадку показник схожості — 0,74, що свідчить про велику вірогідність того, що в ній практично відсутня новизна. У випадку, коли загальна кількість слів статті, що збігається з корпусом, не перевищує чверті загальної кількості слів статті, говорити про показник схожості немає сенсу. Тобто цей числовий метод придатний для аналізу, коли тематичний напрям статті та корпусу збігається.

Висновки. Запропоновану методику визначення новизни можна застосовувати у разі створення репозитаріїв, що сприятиме підвищенню інформаційної якості контенту. Кожну наступну зареєстровану статтю потрібно аналізувати щодо новизни і лише у разі отримання відповідного показника вносити її до репозитарію. В іншому випадку — направляти до рецензента, який приймає остаточне рішення.

Також методи контент-аналізу можна застосовувати для вдосконалення пошукових систем: автоматично проводити маркування [7] однакових інформаційних джерел, щоб не відволікати користувачів на зайвий перегляд "інформаційних близнюків".

Список використаної літератури

1. *Симоненко Т. В.* Мережеве інформаційно-бібліотечне забезпечення наукових досліджень : автореф. дис. … канд. наук із соц. комунікацій / Симоненко Тетяна Василівна ; НАН України, Нац. б-ка України ім. В. І. Вернадського. — К., 2011. — 18 с.
2. *Ландэ Д. В.* Интернетика: навигация в сложных сетях: модели и алгоритмы / Д. В. Ландэ, А. А. Сиарский, И. В. Безсуднов. — М. : Кн. дом "ЛИБРОКОМ", 2009. — 264 с.
3. *Симоненко Т. В.* Репозитарій "Наукова періодика України" як засіб антиплагіату / Т. В. Симоненко // Документознавство. Бібліотекознавство. Інформаційна діяльність: Проблеми науки, освіти, практики : зб. матеріалів VIII Міжнар. наук.-практ. конф., Київ, 17—19 травня 2011 р. — К., 2011. — С. 205—206.
4. *Чернокозинский С. А.* Использование текстовых анализаторов для защиты информации в образовательной среде / С. А. Чернокозинский // Информационное противодействие угрозам терроризма. — 2005. — № 4. — С. 239—242.
5. *Мазов Н. А.* Свободно распространяемые программы для наукометрических и библиометрических исследований // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса : материалы конф. (Судак, АР Крым, Украина 4—12 июня, 2012 г.). — Режим доступа: http://www.gpntb.ru/win/inter-events/crimea2012/disk/123.pdf. — Загл. с экрана.
6. *Пескова О. В.* Методы автоматической классификации документов / О. В. Пескова // Информационные процессы и системы. — 2006. — № 3. — С. 13—20.
7. *Ландэ Д. В.* Поиск знаний в Internet. Профессиональная работа : пер. с англ. / Д. В. Ландэ. — М. : Вильямс, 2005. — 272 с.

Дана методика определения тематической близости и степени новизны публикаций на основе контент-анализа. Статья одного автора сравнивается со статьями другого автора одного тематического направления по частоте использования слов.

A description of content-analysis methods used to determine the originality of research article and their relationships to other articles. Methods include analyzing the frequency of specific words used by one author compared to the frequency of specific words used by other authors who are writing about similar topics.

Надійшла в редакцію 22 червня 2012 року