УДК 004.6

Л. А. Савицька, Т. І. Коробейнікова, В. Д. Тягун

МЕТОД ТА ПРОГРАМНИЙ ЗАСІБ ЗАСТОСУВАННЯ МЕТАДАНИХ В ПРОЦЕСАХ ПОШУКУ

Вінницький національний технічний університет, м. Вінниця

Анотація. Нині дуже популярними є підтримка наукових здобутків науково метричними базами даних. Наука наукометрія дозволяє оцінити глибину науковості статей, дати об'єктивну оцінку новій інформації в тій чи іншій галузі. Наприклад, на відому конференцію ACM Conference on Human Factors in Computer Systems, щороку приходять наукові матеріали на рецензування і кількість їх зросла вдвічі за останні 10 років. Але це лиш ті матеріали, які пройшли рецензування. Якщо ж виміряти ту кількість статей, що поступають на розгляд комісії конференції, то виявимо, що вона виросла в п'ять разів, тобто, реальний приріст інформаційного потоку в тільки одній галузі становить більше ніж в 5 разів за 10 років. Такий стрімкий, майже експоненційний ріст потоку нових наукових розробок означає, що в адекватні часові терміни практично мало хто може встигнути ознайомитись із цими матеріалами, що й призводить до того, що гіпотетично перспективні технології навіть не були помічені ні науковим суспільством в цілому, ні гіпотетичним ринком користувачів. Тому дана робота присвячена розробці та програмній реалізації методу застосування метаданих в процесах пошуку. Цей програмний засіб дозволить досягти збільшення результатів пошуку документів, що задовольняють запиту, в рамках деякої статичної колекції документів.

Високий рівень вирішення поставленої задачі досягнуто за рахунок використання сучасної мови програмування Java. В даній роботі виконано аналіз сучасних моделей, методів та засобів застосування метаданих в процесах пошуку; розглянуто існуючі аналоги та поточний стан технологій в галузі застосування метаданих в процесах пошуку; запропоновано модель та метод застосування метаданих в процесах пошуку; розроблено ключові процеси роботи методу застосування метаданих в процесах пошуку та на його основі цього створено програмний засіб; запропоновано програмну реалізацію запропонованого методу застосування метаданих в процесах пошуку; проведено тестування програмного продукту та виконано аналіз отриманих результатів.

Ключові слова: Метод та програмний засіб застосування метаданих в процесах пошуку, метаданні в процесах пошуку, застосування метаданих в процесах пошуку, метаданні в пошуку.

Abstract. Nowadays, it is very popular to support scientific achievements with scientific metric databases. Scientometric science allows you to assess the depth of scientific articles, to give an objective assessment of new information in a particular field. For example, at the well-known ACM Conference on Human Factors in Computer Systems, every year, scientific papers come for peer review and their numbers have doubled in the last 10 years. But these are only the materials that have been reviewed. If we measure the number of articles submitted to the conference commission, we find that it has increased five times, that is, the real increase in the information flow in only one industry is more than 5 times in 10 years [5]. Such rapid, almost exponential growth in the flow of new scientific development means that in adequate timeframes, almost no one will be able to get acquainted with these materials, which leads to the fact that hypothetically promising technologies have not even been noticed by either the scientific society as a whole or the hypothetical society users. The rapid development of information technology is no longer news, and the further transformation of post-industrial society is closely linked to the acquisition of new knowledge and the effective management of what has already been acquired. Nowadays, it is very popular to support scientific achievements with scientific metric databases. Scientometric science allows you to assess the depth of scientific articles, to give an objective assessment of new information in a particular field. This has been shaping the problem of finding information in specialized systems, archives, including the Internet, for quite some time now. Of course, powerful Google search engines or CiteSeer can make it easier for a person to search and evaluate the relevance of the results they find. However, at the present stage of development, a scientific and creative search is still needed, if not entirely new approaches, it is possible to use non-trivial existing ones, which would improve the situation. Therefore this paper is devoted to the development and software implementation of the method of application of metadata in search processes. This software tool will allow you to increase the search results of documents that meet the query within a static collection of documents. A high level of problem solving has been achieved through the use of modern Java programming language. In this thesis an analysis of modern models, methods and means of using metadata in the search processes; examines the existing analogs and the current state of technology in the application of metadata in the search process; the model and method of application of metadata in search processes is proposed; the key processes of the method of using metadata in the search processes have been developed and, based on this, a software tool was created; the program realization of the proposed method of using metadata in the search processes is proposed; Software product testing was carried out and analysis of the results was performed.

Keywords: Method and software for using metadata in search processes, metadata in search processes, use of metadata in search processes, metadata in search.

Аннотация. Сейчас очень популярны поддержка научных достижений научно метрическими базами данных. Наука наукометрия позволяет оценить глубину научности статей, дать объективную оценку новой информации в той или иной области. Например, на известную конференцию ACM Conference on Human Factors in Computer Systems, ежегодно приходят научные материалы на рецензирование и количество их возросло вдвое за последние 10 лет. Но это лишь те материалы, которые прошли рецензирование. Если же измерить то количество статей, поступающих на рассмотрение комиссии конференции, то обнаружим, что она выросла в пять раз, то есть, реальный прирост информационного потока в только одной отрасли составляет более чем в 5 раз за 10 лет [5]. Такой стремительный, почти экспоненциальный рост потока новых научных разработок означает, что в адекватные временные сроки практически мало кто может успеть ознакомиться с этими материалами, и приводит к тому, что гипотетически перспективные технологии даже не были замечены ни научным обществом в целом, ни гипотетическим рынком пользователей. Поэтому данная работа посвящена разработке и программной реализации метода применения метаданных в процессах поиска. Этот программный средство позволит достичь увеличения результатов поиска документов, удовлетворяющих запросу, в рамках некоторой статической коллекции документов.

Высокий уровень решения поставленной задачи достигнуто за счет использования современного языка программирования Java. В данной работе выполнен анализ современных моделей, методов и средств применения метаданных в процессах поиска; рассмотрены существующие аналоги и текущее состояние технологий в области применения метаданных в процессах поиска; предложена модель и метод применения метаданных в процессах поиска; разработаны ключевые процессы работы метода

применения метаданных в процессах поиска и на его основе этого созданы программное средство; предложено программную реализацию предложенного метода применения метаданных в процессах поиска; проведено тестирование программного про- дукту и выполнен анализ полученных результатов.
Ключевые слова: Метод и программный способ применения метаданных в процессах поиска, метаданные в процессах поиска, применения метаданных в процессах поиска, метаданные в поиске.
DOI: https://doi.org/10.31649/1999-9941-2019-46-3-21-27.

Вступ

Стрімкий розвиток інформаційних технологій вже не новина, подальша трансформація постіндустріального суспільства тісно пов'язана з набуттям нових знань та ефективним керуванням тими, що вже набуті. Такі інтелектуальні прориви є цілком природніми після появи постіндустріального суспільства. Цілком очевидним є те, що підвищення ролі знання веде до стрімкого збільшення інформаційних потоків, які людина отримує щодня. За умов ігнорування частини цього потоку, є ризик втрати актуалізації своїх знань. І навпаки, за умов вмілого та ефективного керування потоками інформації дозволяє зберігати та підвищувати професійну кваліфікацію, просуватися вперед у щоденних справах, створювати новий контент тощо. Закономірно, що рано чи пізно, на якомусь етапі інформаційний потік може стати настільки величезним, що навіть при великому бажанні, і витрачаючи на освоєння знань більшу частину свого часу, людина не встигатиме – бо новий контент буде з'являтись швидше, ніж попередній може бути опрацьованим в принципі. Часто нові технології так і не впроваджуються у життя, чи не виходять на ринок, оскільки встигають морально застаріти, і головною причиною такого стану справ було те, що системи пошуку не встигли їх з різних причин адекватно проіндексувати і ці матеріали так і залишилися незнайденими в безкінечних множинах інформаційних потоків.

Актуальність

З метою полегшити процеси пошуку і виключити аналіз всього тексту в пошуках ключових слів та виразів, в даній роботі пропонується використовувати так звані «мета-теги», які в стислій формі описують вміст сторінки, текстових даних чи якогось ресурсу. Таким чином, кожна веб-сторінка, чи документ, супроводжуються «хмарою тематичних тегів». Звісно їх кількість, на жаль, кінцева, а із швидким ростом інформаційного потоку, на кожен такий тег припадатиме все більше і більше даних, а це, відповідно, вимагатиме додаткових уточнюючих тегів, і зрештою, кожен документ треба буде супроводжувати такою кількістю тегів, що їх індексація за своїм обсягом не буде відрізнятись від повнотекстової інформації.

Мета

Метою дослідження статті є збільшення результатів пошуку документів, що задовольняють запиту, в рамках деякої статичної колекції документів.

Для досягнення поставленої у роботі мети необхідно розв'язати такі завдання:

- провести аналіз сучасних моделей, методів та засобів застосування метаданих в процесах пошуку;
- розглянути існуючі аналоги та поточний стан технологій в галузі застосування метаданих в процесах пошуку;
- запропонувати модель та метод застосування метаданих в процесах пошуку;
- розробити ключові процеси роботи методу застосування метаданих в процесах пошуку та на його основі цього розробити програмний засіб;
- виконати програмну реалізацію запропонованого методу застосування метаданих в процесах пошуку;
- провести тестування програмного продукту та виконати аналіз отриманих результатів.

Методи застосування метаданих в процесах пошуку

У загальному випадку, визначень та класифікацій для широкого поняття метаданих існує немало. Часто, в загальнодоступних джерелах, використовується такий комплект критеріїв, що є спільним для всіх їх типів. Розглянемо для прикладу звичайний текст. Отже, його метадані можуть містити таке:

1) Перелік засобів, що використані і за допомогою яких створено даний текст;
2) Мета цього конкретного тексту;
3) Час та дата створення цього конкретного тексту;
4) Автор цього конкретного тексту;
5) Розташування комп'ютера (гаджета, вузла, хоста), на якому він був створений (розміщений, модифікований тощо);
6) Використані стандарти для відображення цього конкретного тексту.

Часто ці метадані зберігались і знаходилися разом із самим текстом, що вони його характеризують, тобто, в складі формату самого документу. Так, скажімо, відповідні поля для цих дій є в форматах фірми Microsoft, а також у багатьох інших, можливо менш поширених способах зберігання інформації та даних.

Більше того, подібні поля присутні також в графічних форматах (EXIF, JFIF) та, звісно, у звукових форматах (у вигляді т.з. ID3-тегів).

Проте з часом лише цього стало мало, бо цих даних досить для автоматичної категоризації, як вже говорилось вище, але недосить для здійснення складного процесу пошуку і гарної оцінки релевантності текстів. Тому американський фахівець по базам даних (БД) Ральф Кімбелл запропонував розширити систему метаданих загалом, і припустив, що можна поділити матедані на два потужних класи:

- технічні метадані;
- бізнес-метадані.

Моделі та технології застосування метаданих у процесах пошуку в Інтернеті

Семантичні структури можуть бути реалізовані досить різними способами. Це може бути реалізовано за допомогою структур, що знаходяться усередині самого файлу, або ж це може бути якась окремо створена спеціальна система, призначена для збирання і зберігання саме семантичних даних.

Одною із перших таких спроб створити подібну систему запропонували Рой Голдман та Дженніфер Уідом. Їхня пропозиція зрештою зводилась до необхідності створення такої особливої БД, до створення т.з. «сітки даних», яка би містила форматовані нотатки та довідки про інформаційні матеріали і ресурси, і які би згодом стали основою для формулювання все більш точних запитів та збирання все більшої кількості статистики, що би дозволила оптимізувати процес пошуку. Саме цими авторами було вперше запропоновано застосувати відому деревоподібну модель деталізації метаданих, та ними ж були сформовані базові поняття, на яких створено сучасні семантичні системи. Нехай існує БД (на прикладі типів закладів харчування) із такою структурою (рис. 1).

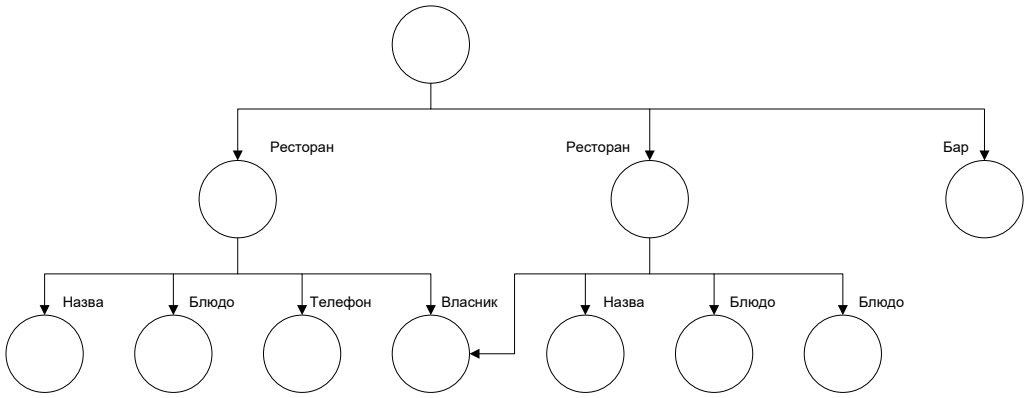

Рисунок 1 – Деревоподібна модель БД із нечіткою структурою метаданих

Як видно з рис. 1, в цій БД однакові об'єкти, як не дивно, можуть мати різний перелік властивостей. Причому, деякі із них можуть навіть повторюватись, а деякі – можуть загалі бути відсутніми. З метою впорядкувати все це і дозволити користувачу системи працювати однотипно із усіма елементами системи, незважаючи на їх відмінності, дана БД використовує механізм метаданих, організованих у відповідності до видів і типів елементів. Це може бути окрема структура, що містить список можливих властивостей, які може мати окремий елемент (рис. 2).

Подібний підхід запропонували запровадити і для текстових даних. Згідно цього, в текстовому документі якимось чином виділялися певні «поля» (скажімо, маркувались спеціальними символами, або спеціальними методами розмітки тексту).

Отож, вже нема такої необхідності переглядати весь текст повністю, а можна лиш перевіряти його окремі ключові теги, і ще причому лиш ті, які вказав користувач. Тепер, в поєднанні з процесоми індексування та хешування цих полів, ми можемо значно пришвидшувати пошук, і довести повнотекстовий пошук до тієї ж швидкості, що й БД.

Модель застосування метаданих в процесах пошуку

Тегування електронних записів, в тому числі і нотаток, що актуально для теми даної магістерської роботи, використовується з метою полегшення процесів пошуку через можливість категоризувати дані всередині електронних записів. Кожному із записів ставиться у відповідність комплект ключових слів (тегів), кожне із яких потім індексується, і отож користувач може переглядати та групувати свої електронні закладки з точки зору різних тематик, чи у відповідності певним задачам. Система тегів була згодом розвинута в модель соціального тегування, яка була покладена в основу Web 2.0.

Першопочатково модель застосування метаданих в процесах пошуку була створена для пришвидшення людського пошуку. Організація контенту за тегами і ключовими словами дозволяла людям самостійно утворювати ієрархії категорій, і отож брати на себе частку функцій та впливати на процеси застосування метаданих в процесах пошуку.

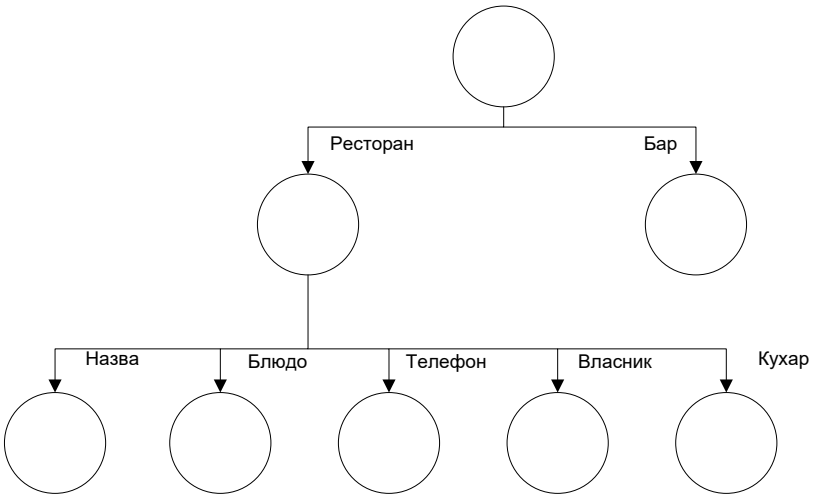

Рисунок 2 – Деревоподібна модель БД із чіткою структурою метаданих

Цілком справедливим є факт, що кількість тегів із зростанням кількості матеріалу завжди також росте, причому зростає експоненційно. Тоді виникає т.з. «хмара тегів», яку часто ілюструють діаграмою частотного розподілу їх використання. Фактично «хмара» є видом зваженого списку, де всі теги групуються відповідно їх частотному розподілу. Графічним і візуальним чином це можна показати на рис. 3, де частота позначається розміром шрифта.

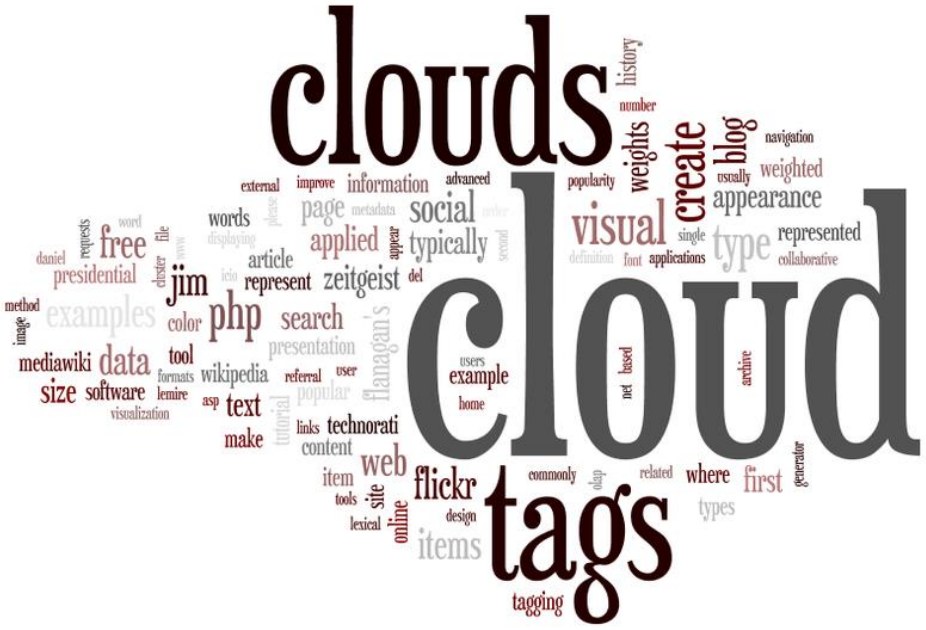

Рисунок 3 – Приклад хмари тегів

Далі такий зважений список групується за певними категоріями, далі утворюється так звана «глобальна хмара метаданих» із кластерами, що поділені за елементами та користувачами. Утворені кластери можуть сортуватися за кількістю підкатегорій.

Нечіткий пошук в такій системі із застосуванням метаданих зводиться до гіпотези, що у випадку досить простих тегів вони співпадатимуть із ключовими словами, які задає людина в рядку пошуку, і сам процес пошуку відтак стане більш ефективним. На практиці ця гіпотеза цілковито не потвердилась.

Приична в тому, що окрема особа вибудовує власну термінологію ключових слів, керуючись особистою лексикою, яка часто не має відношення до граматичних чи семантичних правил.

Скажімо, синонімічний ряд тегів для ОС Linux виглядає так: Linux, GNU/Linux, Линакс, Лінукс, Лінуха, Лялікс, Лунікс. Це не враховуючи різновидів цієї ОС, які теж вживаються як частина синонімічного ряду. Все це формує явище «фолксономії» (рис. 4).

Розглянемо приклад. Нехай користувач створив деякий електронний матеріал, розмістив на якомусь реусрсі, і маркував 5 тегами, які відображають різні аспекти матеріалу.

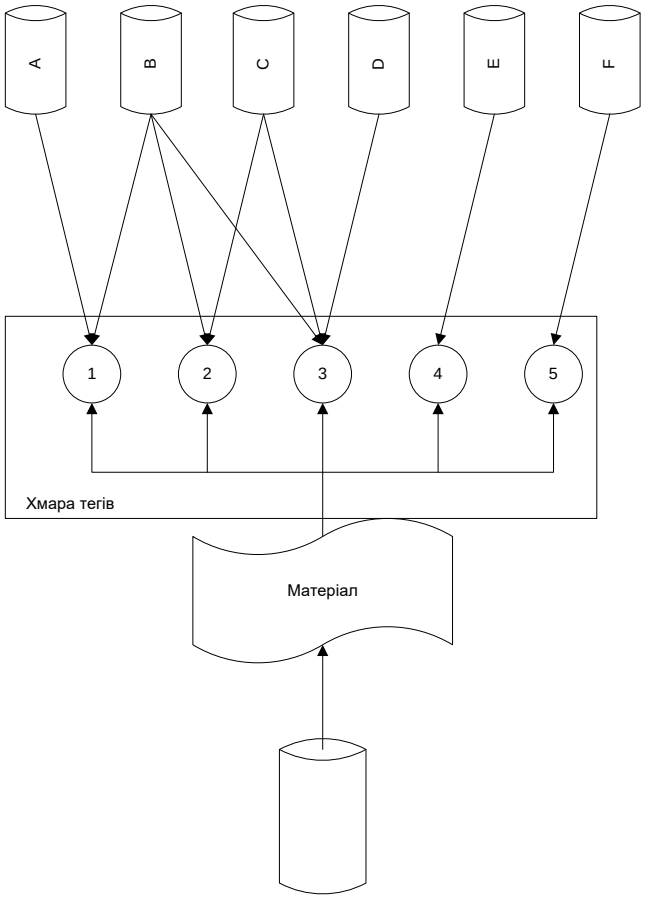

Рисунок 4 – Широке фолксономічне тегування

Під час подальшого поширення цього матеріалу, він може бути тегований не всіма цими ключовими словами, а лиш тими, які будуть такими для інших користувачів. І тоді може бути так, що у результаті цей матеріал ніби-то «зникне» із індексів, які відповідають аспектам 4 і 5 (бо вони лишаться в поширеннях значно меншої кількості користувачів), і тоді цей матеріал буде категоризований із більшою ймовірністю в кластери із застосуванням метаданих, що відповідають аспектам 1, 2 і 3 (умовно). І тоді цей матеріал взагалі ризикує бути «не знайденим», бо логічний ланцюжок, який пов'язує всі 5 маркерів матеріалу, будуть втрачені.

Такий фолксономічний підхід дозволяє користувачам систем пошуку формувати свої кластери понять і визначеннь, і навіть свою власну хмару мета-даних. Проте при цьому дуже важко знайти об'єкти, які носять загальний чи не дуже окремий характер. Виникає така собі плутанина, і тоді знайти цей матеріал можна, звернувшись до самого автора, або досить довго заглибившись в його публікації. Звісно, що на це в користувача нема часу, і виникає та сама задача, яка була поставлена ще у вступі до даної роботи – проблема т.з. «незнаходжуваності» матеріалу.

Вирішити цю задачу, можна, зберігаючи на якомусь ресурсі ці логічні ланцюжки, і саме тут і може бути корисною модель застосування метаданих в тегах в процесах пошуку в Інтернет.

Вже згадувалося, що за допомогою мета-сутностей можна будувати ієрархії визначень та їх логічних взаємозв'язків. Є ідея дозволити користувачу створювати свої мета-сутності. Тоді, під час процесів пошуку, інший користувач (або пошукова система в автоматичному режимі) зможе звернутись до мета-сутності автора матеріалу, і таки мчином легко відновити всі необхідні логічні зв'язки. Ступінь

знаходжуваності документа при такому підході значно підвищується. Звісно, що зільшиться обсяг самих метаданих, але це не становить вели коїпроблеми із сучасним розвитком обчислювальної та комп'ютерної техніки, де обсяги пам'яті уже не має такого критичного значення, як це було іще 10 років назад.

Підхід цей, використовується в «фасетній класифікації», що полягала в узгодженні різних бібліотечних класифікацій, які існували в світі, і створенні «сітчастої» моделі класифікації, де кожна класифікаційна система – це вісь, яку перетинають інші класифікації в точках, де їх визначення співпадали. Звісно, що сама така сітчаста структура не може точно категоризувати матеріал, проте вона дозволяє переходити від однієї класифікації до іншої, спираючись на спільні визначення. Такий метод узгодження різних ієрархій визначенньь називається багатоаспектним класифікуванням.

Саме цю модель поклали в основу застосування метаданих до процесів пошуку, і саме наведені елементи фасетної класифікації можна застосувати для реалізації нечіткого пошуку по користувацьким тегам.

Висновки

Результатом статті є збільшення результатів пошуку документів, що задовольняють запиту, в рамках деякої статичної колекції документів. Було спроектовано демонстраційний приклад, який дозволяє показати, яким чином можна здійснювати нечіткий пошук на основі хмари тегів та семантичних мета-сутностей.

Наукова новизна одержаних результатів полягає в такому:

- вперше запропоновано метод застосування метаданих в процесах пошуку. Запропонований у даній роботі метод застосування метаданих в процесах пошуку дозволяє збільшити результатів пошуку документів, що задовольняють запиту, в рамках деякої статичної колекції документів;
- вдосконалено модель застосування метаданих в процесах пошуку, що дозволяє створення користувачем власних мета-сутностей;
- вдосконалено процес роботи з мета-даними, який дозволяє створення семантичних ланцюжків між поняттями;
- вдосконалено процес створення мета-сутностей, які дозволяють зменшувати кількість тегів у хмарі тегів.

Запропонований метод застосування метаданих в процесах пошуку, що реалізований у демонстраційній моделі в інтегральному показнику дозволяє збільшити результатів пошук документів, що задовольняють запиту, в рамках деякої статичної колекції документів на 12,7% у порівнянні із аналогами.

Список літератури

[1] Э. Тоффлер Третья волна – М., АСТ, 2016.
[2] V. Bush As We May Think – The Athlantic Monthly, 1 July, 2005 - http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/
[3] A. Kay The Early History of Smalltalk – ACM, 2013 - http://www.smalltalk.org/smalltalk/TheEarlyHistoryOfSmalltalk_TOC.
[4] Хороший план №9 // Хакер, 29.07.2014 - http://www.xakep.ru/post/23246/default.asp
[5] C. Bartneck, J. Hu Scientometric Analysis of The CHI Proceedings - Eindhoven University of Technology, Dept. of Industrial Design, - April, 2019
[6] S. Brin, L. Page The Anatomy of a Large-Scale Hypertextual Web Search Engine – Stanford University, Computer Science Dept., 2018
[7] H. Li, I. Councill, L. Bolelli, D. Zhou, Y. Song, and others CiteSeer – A Scalable Autonomous Scientific Digital Library – Pennsylvania State University, 2017.

Стаття надійшла: 11.11.2019.

References

[1] E. Toffler Tretya volna – M., AST, 2016.
[2] V. Bush As We May Think – The Athlantic Monthly, 1 July, 2005 - http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/
[3] A. Kay The Early History of Smalltalk – ACM, 2013 - http://www.smalltalk.org/smalltalk/TheEarlyHistoryOfSmalltalk_TOC.
[4] Horoshiy plan #9 // Haker, 29.07.2014 - http://www.xakep.ru/post/23246/default.asp
[5] C. Bartneck, J. Hu Scientometric Analysis of The CHI Proceedings - Eindhoven University of Technology, Dept. of Industrial Design, - April, 2019
[6] S. Brin, L. Page The Anatomy of a Large-Scale Hypertextual Web Search Engine – Stanford University, Computer Science Dept., 2018

[7] H. Li, I. Councill, L. Bolelli, D. Zhou, Y. Song, and others CiteSeer – A Scalable Autonomous Scientific Digital Library – Pennsylvania State University, 2017.

Відомості про авторів

Коробейнікова Тетяна Іванівна, к. т. н., доцент кафедри обчислювальної техніки, ВНТУ, кафедра обчислювальної техніки, Вінниця, Хмельницьке шосе, 95.

Савицька Людмила Анатоліївна, к. т. н., доцент кафедри обчислювальної техніки, ВНТУ, кафедра обчислювальної техніки, Вінниця, Хмельницьке шосе, 95.

Тягун Дмитро Тарасович, магістр кафедри обчислювальної техніки, ВНТУ, кафедра обчислювальної техніки, Вінниця, Хмельницьке шосе, 95.

Л. А. Савицкая, Т. И. Коробейникова, Д. Т. Тягун

МЕТОД И ПРОГРАММНОЕ СРЕДСТВО ПРИМЕНЕНИЯ МЕТАДАННЫХ В ПРОЦЕССЕ ПОИСКА

Винницкий национальный технический университет, г. Винница

L. A. Savytska, T. I. Korobeinikova, D. T. Tyagun

METHOD AND SOFTWARE FOR APPLYING METADATA TO SEARCH PROCESSES

Vinnitsa National Technical University, Vinnitsa