

Ранжирование документов в системе поиска, основанной на применении онтологии

© В.Т. Вдовицын

© В.А. Лебедев

Институт прикладных математических исследований Карельского научного центра РАН
Петрозаводск
vdov@krc.karelia.ru

Аннотация

В работе предлагается алгоритм ранжирования документов, разработанный в плане развития технологии систематизации и поиска информации с применением онтологии. При этом учитываются как специфические особенности самой технологии, так и некоторые известные приемы построения функций ранжирования для систем поиска, основанных на применении ключевых слов.

1 Введение

Разработка эффективных систем поиска в огромных массивах слабоструктурированных документов остается актуальной проблемой. Традиционные поисковые системы, которые позволяют пользователю выразить свои информационные потребности путем задания списка ключевых слов, обладают рядом существенных недостатков. К их числу следует отнести трудности, связанные в первую очередь с многозначностью ключевых слов, недостаточным знанием терминологии предметной области, а также сложности формулирования запросов с использованием булевских операторов [1]. С другой стороны, очень часто поисковая система выдает по запросу пользователя большой массив релевантных запросу документов и далеко не всегда ранжирует их в соответствии с информационными потребностями пользователя.

Во многих поисковых системах используемые методы ранжирования документов, учитывающие так называемые страничные факторы, базируются на применении статистической информации о распределении ключевых слов в запросе и текстах документов. Существуют различные подходы к построению функций ранжирования, например, в системах поиска, базирующихся на векторной модели, проблема построения функции ранжирования документов в основном сводится к задаче определения весовых коэффициентов терминов.

В данной работе предлагается подход к построению алгоритма ранжирования документов в системе

поиска, основанной на применении онтологии. При этом учитываются как специфические особенности самой технологии, так и некоторые известные приемы построения функций ранжирования [2].

2 Технология систематизации и поиска документов, основанная на применении онтологии

Одним из перспективных направлений исследований и разработок, направленных на повышение эффективности информационного поиска, является применение онтологий (ontology-based information retrieval). Такие системы информационного поиска учитывают смысловое содержание терминов запроса, используют онтологии, как для индексации информационных ресурсов, так и для организации семантического поиска в больших массивах документов. При этом исследуется и решается ряд проблем, связанных, например, с тем, какими преимуществами обладают системы информационного поиска основанные на применении онтологии по сравнению с традиционными системами, осуществляющими поиск по ключевым словам? Как выразить информационные потребности пользователя на онтологически-ориентированном языке (например, RDQL)? Исследуются также проблемы «неоднозначности» терминов, адаптации векторной модели поиска к особенностям онтологически-ориентированного поиска, вопросы ранжирования найденных по запросу документов и др. Необходимо отметить, что проведенные многими исследователями эксперименты по оценке эффективности онтологически-ориентированных систем поиска (по критерию – «точность/полнота») демонстрируют их преимущества по сравнению с системами поиска по ключевым словам [3–5, 7].

В течение ряда последних лет нами разрабатывается и исследуется онтологически-ориентированная технология систематизации и поиска электронных публикаций [7–10]. При этом под онтологией понимается «формальное представление множества понятий предметной области и связей между этими понятиями» [11].

В основу построенной онтологии положены: рубрикатор (в нашем случае ГРНТИ); набор логических условий предметизации документов (для их распределения по соответствующим рубрикам ГРНТИ); таксономия терминов определенной

научной предметной области, термины которой связаны отношениями классификации «род–вид», агрегации, «часть–целое» и синонимии.

Процесс систематизации публикаций разделяется на два этапа: предметизацию и индексацию. В качестве информационной основы предметизации используются термины таксономии и набор логических условий (логических функций, описывающих связи научных терминов по определенной тематике исследований), с помощью которых осуществляется процесс отнесения публикаций к соответствующим рубрикам (в нашем случае – к рубрикам ГРНТИ). Для формирования этих логических условий, описывающих содержание публикаций, используются термины таксономии и логические операторы: AND, OR, NOT.

Ниже приведен пример логического условия предметизации, представленного в виде простого правила-продукции экспертной системы.

IF (охлаждение **OR** температура **OR** влага **OR** влажность **OR** нестабильный климат **OR** устойчивость **OR** стойкость **OR** выживаемость **OR** адаптация **OR** терморезистентность **OR** реакция) **AND** (растения **OR** пшеница **OR** картофель)

THEN рубрика ГРНТИ – 34.31.15. Действие физических факторов на растения

В настоящее время сформулирован ряд логических условий (правил-продукций) для предметизации публикаций по некоторым направлениям биологии, почвоведения, лесному хозяйству и водным ресурсам, относящихся к направлениям научных исследований институтов КарНЦ РАН.

Процесс индексации состоит из двух основных этапов. На первом этапе выполняется нормализация текста – каждая публикация переводится из формата PDF в формат TXT, из текста удаляются «малоинформативные» слова, к тексту и терминам выделенного фрагмента таксономии применяется алгоритм стемминга (в нашем случае – Стеммер Портера). На втором этапе осуществляется последовательное сканирование текста публикации и сопоставление каждого слова с терминами выделенного фрагмента таксономии, характеризующего содержание предметной рубрики. Т.е. в процессе индексации последовательно обходятся поддерева всех рубрик, к которым была отнесена публикация на этапе предметизации. При этом каждый раз производится поиск термина таксономии в тексте публикации и если обнаруживается такое вхождение термина, то индексируется не только этот термин, но и все его предки из исследуемого поддерева рубрики. Таким образом, индекс публикации представляет собой упорядоченную совокупность терминов таксономии, и на наш взгляд более детально характеризует ее содержание по сравнению с традиционным списком ключевых слов.

Таксономия и база индексов публикаций обеспечивают тематический поиск публикаций по запросам пользователей. Нами разработана технология построения и исполнения запросов, суть

которой заключается в следующем. Пользователю сначала предлагается выбрать рубрику ГРНТИ, которая, по его мнению, должна содержать интересующие его материалы (если этих рубрик не одна, то придется построить несколько однотипных запросов). Далее ему предлагается соответствующий рубрике фрагмент таксономии, в котором он должен отметить интересующие его термины. С использованием указанных терминов система автоматически формирует запрос в виде логического выражения, определяющего конъюнктивные и/или дизъюнктивные связи терминов.

Следует отметить, что поскольку поиск по запросу осуществляется в базе индексов (а не в текстах электронных публикаций), запрос автоматически расширяется включением в него конъюнкции терминов от корня и дизъюнкции терминов и их синонимов вплоть до листьев от указанных пользователем терминов. Тем самым обеспечивается повышение точности ответа на запрос за счет конъюнкции терминов предыдущих уровней таксономии и полноты за счет дизъюнкции терминов одного уровня таксономии и их синонимов. Список названий найденных по запросу публикаций выводится пользователю в виде гиперссылок для последующего просмотра или сохранения текстов публикаций в «личном» кабинете пользователя (рис.1).

Таким образом, использование базы индексов как результата систематизации публикаций непосредственно для их поиска обеспечивает с одной стороны устранение полисемии терминов (т.е. устраняет многозначность терминов за счет «отсечения» других предметных областей в процессе построения запроса), а с другой определяет конкретную предметную область запроса. Тем самым обеспечивается как релевантность, так и пертинентность найденных системой по запросу документов.

Следует также заметить, что пользователю на наш взгляд гораздо проще и точнее выразить свои информационные потребности путем указания терминов в таксономии по сравнению с заданием списка ключевых слов. При этом ему не надо формировать логические условия отбора данных с использованием логических операторов: AND, OR, NOT (система делает это автоматически).

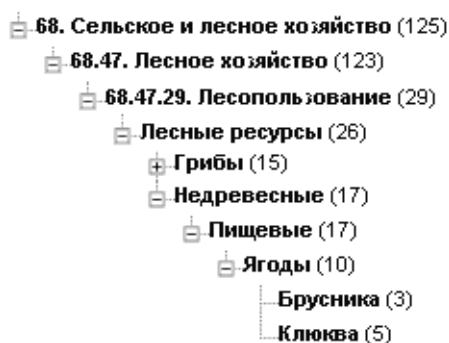
Апробация разработанной технологии проводится в рамках создания и развития информационно-аналитической системы «Природные ресурсы Карелии» – <http://ias.krc.karelia.ru> [12].

На основе разработанной онтологически-ориентированной технологии систематизации и поиска информации можно построить ряд информационных систем различного функционала и назначения. Например, может быть разработана информационная система для оперативной (в режиме on-line) поддержки деятельности спортивных журналистов. Для этого потребуется разработать соответствующую предметную онтологию (рубрикатор видов спорта, ряд логических условий пред-

Поиск по таксономии терминов

выделить все

убрать выделение



Для ключевых слов: 68. Сельское и лесное хозяйство И 68.47.

Лесное хозяйство И 68.47.29. Лесопользование И Лесные ресурсы И (Грибы) ИЛИ (Недревесные И Пищевые И Ягоды И (Брусника) ИЛИ (Клюква))

найдено 17 документов

Документы коллекции "Электронные версии научных публикаций, изданных в РИО КарНЦ РАН"

1. Афиллофоровые грибы заповедника "Кивач"
2. Вепское и прибалтийско-финское в некоторых наименованиях ягод (на материале ALFE)
3. Грибы
4. Исследования по биоте афиллофороидных грибов в таежных экосистемах Северо-Запада России
5. Лесные ресурсы таежной зоны России: проблемы лесопользования и лесовосстановления: Материалы Всеросс. науч. конф. с международ. участием (Петрозаводск 30.09-03.10.2009 г.)
6. Лесные съедобные грибы и их использование в Карелии
7. Методические подходы к капитализации лесных ресурсов региона
8. Микроиндикационная методика в лесопромышленном комплексе Республики Карелия
9. Научные разработки Института леса КарНЦ РАН и их реализация в области лесопользования и лесовосстановления
10. О проблемах использования болот в лесном фонде
11. О распространении и охранном статусе видов афиллофоровых грибов, включенных в Красную книгу Республики Карелия
12. Освоение заболоченных лесов как фактор интенсификации лесопользования в Республике Коми
13. Основы лесного хозяйства для лесопользователей

Рис. 1. Поиск по онтологии, соответствующий рубрике «Сельское и лесное хозяйство».

метизации для распределения поступающих новостей по рубрикам, таксономию терминов предметной области и их синонимов), а также сформировать «персональный профиль» пользователя–журналиста, определяющий его информационные потребности.

В данной работе рассматривается модель информационной системы, предназначенной для систематизации, поиска и ранжирования электронных научных публикаций, соответствующих информационным потребностям пользователя. Предполагается, что свои информационные потребности такой пользователь выражает путем указания в таксономии терминов определенной научной предметной области соответствующих терминов. На их основе система автоматически сформирует его «персональный профиль», который будет использоваться для систематизации, поиска и ранжирования электронных научных публикаций. Массив публикаций может регулярно пополняться (например, с помощью тематического краулера), а система в автоматическом режиме будет пополнять

«личный кабинет» пользователя новыми найденными публикациями, релевантными его информационным потребностям в соответствии с заданным персональным профилем пользователя (при этом все найденные новые публикации получают помету NEW, а по электронной почте пользователю могут приходиться уведомительные сообщения).

3 Алгоритм ранжирования документов, основанный на применении онтологии

Первоначально мы разрабатывали схему ранжирования документов, основываясь на традиционном (статистическом) подходе. В общем виде такая схема ранжирования выглядит следующим образом. Первый ранг назначался документам, в которых полный набор терминов запроса входит в его название и аннотацию. Далее определялась частота вхождения набора терминов запроса в тексте документов, и вычислялось отношение этого числа к числу страниц текста. Если это отношение было не меньше половины, то

документу присваивался второй ранг, а если это отношение получалось меньше 0.5 – третий ранг. После чего выполнялось упорядочивание документов, полученных системой при формировании ответа на запрос, в соответствии с назначенными рангами.

В основу модифицированного алгоритма ранжирования документов положено предположение о том, что указанные пользователем (при формировании запроса или задании персонального профиля) в таксономии термины, расположенные на «нижних» уровнях древовидной структуры (представляющей таксономию), в большей степени определяют для него «ценность» публикации, чем термины, расположенные на «верхних» уровнях этого дерева. Также мы учитываем и тот факт, что «ценность» публикации для пользователя во многом определяется и тем, в какой зоне текста публикации наиболее часто появляется термины запроса (например, в научных статьях можно выделить следующие зоны: название, ключевые слова, аннотация, основной текст и т.п.). Если термин запроса появляется в названии и/или в списке ключевых слов, то можно предположить, что эта публикация в большей степени соответствует информационным потребностям пользователя, чем иные публикации, в которых этого не зафиксировано (аналогичное предположение учитывается, например, в алгоритме OKAPI BM25F [13]).

С учетом этих, на наш взгляд вполне разумных предположений, предлагается модифицированный алгоритм ранжирования, разработанный для рассматриваемой модели информационной системы, который можно представить, в самом общем виде, следующим образом.

По сформулированному запросу (или по заданному профилю пользователя) формируется расширенный вектор терминов $T = (T_n, T_{n-1}, \dots, T_1)$, где: T_1 – корневой термин выделенного фрагмента таксономии, а T_n – термин, расположенный на «концевой» вершине дерева (представляющего таксономию терминов) и лежащий на соответствующем пути дерева от «последнего» указанного в запросе термина (промежуточные термины T_{n-1}, \dots, T_2 составляют путь в дереве). Таких векторов, сформированных по запросу, может быть несколько, и все они упорядочиваются по убыванию длины. При этом каждый такой вектор определяет название раздела, в который будут помещаться найденные системой по запросу публикации.

Предполагается, что в результате выполнения запроса все найденные публикации распределяются по соответствующим разделам (наименование раздела формируется из списка терминов, указанных в векторе T), а внутри каждого раздела все публикации сортируются по значению их весов, вычисленных с помощью предложенной функции ранжирования.

Для построения функции ранжирования введем следующие обозначения:

$T = (T_n, T_{n-1}, \dots, T_1)$ – расширенный вектор терминов (таких векторов может быть несколько, все они упорядочиваются по длине, чем «длиннее» вектор, тем «ценнее» должны быть найденные системой на основе данных терминов публикации);

$(v_n, v_{n-1}, \dots, v_1)$ – веса компонент вектора T ($v_n > v_{n-1} > \dots > v_1$, значения весов можно вычислить, например, по следующему правилу: $v_i = \log_2 10^i$, $i = 1, 2, \dots, n$);

$(\psi_1, \psi_2, \dots, \psi_m)$ – веса, приписанные определенным зонам публикации, куда могут входить термины запроса (например, если мы учитываем вхождение термина в название публикации, список ключевых слов, аннотацию, основной текст публикации, то в этом случае $m = 4$). Для научных публикаций логично предположить, что $\psi_1 > \psi_2 > \psi_3 > \psi_4$ (т.е. термины запроса, входящие в название публикации, являются более значимыми при ранжировании);

$\{x_{ji}\}$ – число вхождений данного термина в соответствующее поле публикации, $j = 1, \dots, m$; $i = 1, \dots, n$;

w – вес публикации, вычисленный с помощью функции ранжирования.

Тогда, функция ранжирования для оценки веса публикаций в нашем случае будет иметь следующий вид:

$$w = \sum_{i=1}^n v_i * \sum_{j=1}^4 \psi_j * x_{ji} \quad (1)$$

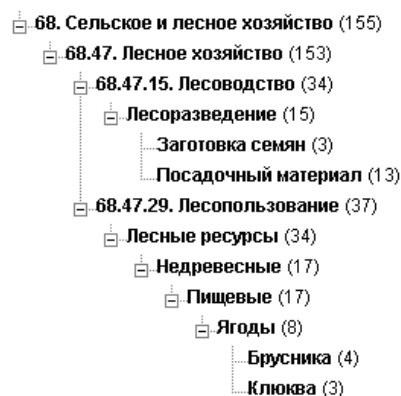
На рис.2 приведен пример запроса, результаты выполнения которого упорядочены в соответствии с предложенным в работе модифицированным алгоритмом ранжирования публикаций.

Следует отметить, что в функции ранжирования (1) учитывается встречаемость терминов запроса в соответствующих зонах публикации (например, в нашем случае учитывается число вхождений термина запроса в следующие выделенные зоны: название публикации, список ключевых слов, аннотацию и основной текст). Если в первых трех случаях достаточно использовать количество вхождений каждого термина в соответствующую зону публикации (таких вхождений будет немного – 1 или 2), то количество вхождений термина в основной текст публикации зависит от размера конкретной публикации («большая» публикация может содержать больше повторений одного и того же термина, чем «маленькая», но, тем не менее, являться более релевантной запросу пользователя). В таких случаях обычно используют параметр «вес термина», который обозначим через ω_{4i} (вес термина i в 4 зоне – в тексте публикации) и определим его следующим образом: $\omega_{4i} = 1 + \log_{10} x_{4i}$, если $x_{4i} > 0$; и 0 – в противном случае. С учетом этого формула (1) будет выглядеть следующим образом:

$$w = \sum_{i=1}^n v_i * (\sum_{j=1}^3 \psi_j * x_{ji} + \psi_4 * \omega_{4i}) \quad (2)$$

Поиск по таксономии терминов

выделить все | убрать выделение



Для ключевых слов: 68. Сельское и лесное хозяйство **И** 68.47. Лесное хозяйство **И** 68.47.29. Лесопользование **И** Лесные ресурсы **И** Недревесные **И** Пищевые **И** Ягоды **И** Клюква (8)

1. Лесные ресурсы таежной зоны России: проблемы лесопользования и лесовосстановления: Материалы Всеросс. науч. конф. с международ. участием (Петрозаводск 30.09-03.10.2009 г.) (205)
2. 1.4. Ягодные растения болот (182)
3. Вепское и прибалтийско-финское в некоторых наименованиях ягод (на материале ALFE) (103)

Для ключевых слов: 68. Сельское и лесное хозяйство **И** 68.47. Лесное хозяйство **И** 68.47.29. Лесопользование **И** Лесные ресурсы **И** Недревесные **И** Пищевые **И** Ягоды **И** Брусника (8)

1. Лесные ресурсы таежной зоны России: проблемы лесопользования и лесовосстановления: Материалы Всеросс. науч. конф. с международ. участием (Петрозаводск 30.09-03.10.2009 г.) (216)
2. 1.4. Ягодные растения болот (128)
3. Районирование таежных лесов по ресурсным, хозяйственным и экологическим параметрам на ландшафтной основе (99)
4. Вепское и прибалтийско-финское в некоторых наименованиях ягод (на материале ALFE) (98)

Для ключевых слов: 68. Сельское и лесное хозяйство **И** 68.47. Лесное хозяйство **И** 68.47.15. Лесоводство **И** Лесоразведение **И** Посадочный материал (5)

1. Лесовосстановление на вырубках Северо-Запада России (87)
2. Лесные ресурсы таежной зоны России: проблемы лесопользования и лесовосстановления: Материалы Всеросс. науч. конф. с международ. участием (Петрозаводск 30.09-03.10.2009 г.) (76)
3. Книга юного лесовода: Учебное пособие по основам лесоведения, лесоводства и охраны природы для обучающихся по дополнительным образовательным программам (75)
4. Основы лесного хозяйства для лесопользователей (70)
5. Продуктивность и устойчивость лесных почв: Материалы III Международ. конф. (Петрозаводск 7-11.09.2009 г.) (38)

Другие документы...

Показать все результаты

Рис. 2. Пример поиска по онтологии, соответствующий рубрике «Сельское и лесное хозяйство», с ранжированием найденных публикаций по предложенному алгоритму.

Таким образом, процедура ранжирования публикаций в онтологически-ориентированной системе поиска состоит в следующем. Во-первых, все найденные по запросу публикации распределяются системой по разделам (при этом наименование каждого раздела формируется из соответствующего расширенного списка терминов запроса). Во-вторых, разделы упорядочиваются в соответствии с длиной соответствующего вектора T. В-третьих, внутри каждого раздела все найденные публикации упорядочиваются в соответствии с их весами, вычисленными с помощью построенной функции ранжирования (2). Кроме того, в функцию ранжирования можно включить и дополнительные параметры (например, индекс цитирования публикации, количество обращений к данной публикации и т.п.), которые также могут служить основанием для их первоочередного просмотра.

Для оценки эффективности предложенного метода ранжирования нами запланированы и проводится серия вычислительных экспериментов и сравнение результатов работы нашей системы с результатами поисковой системы «Яндекс. Персональный поиск». В системе персонального поиска «Яндекс» реализованы свои механизмы полнотекстового поиска документов по ключевым словам и ранжирование полученных результатов, детали которых нам не известны. Для проведения вычислительного эксперимента были отобраны электронные публикации, которые на этапе предметизации были отнесены нашей системой к рубрикам «68.47.29. Лесопользование» (36 статей) и «68.47.15. Лесоводство» (39 статей). При этом в предложенной формуле ранжирования (2) учитывались значения весов терминов, входящих в название документа и в его основной текст (т.е. значения весов терминов запроса, входящих в аннотацию и список ключевых слов публикации на

Название публикации (в скобках указан вес публикации, вычисленный с помощью предложенного алгоритма ранжирования)	Ранг публикации	
	Поиск по таксономии	Яндекс
Лесные ресурсы таежной зоны России: проблемы лесопользования и лесовосстановления: Материалы Всеросс. науч. конф. с междунар. участием (Петрозаводск 30.09-03.10.2009 г.) (66)	1	3
Основы лесного хозяйства для лесопользователей (54)	2	1
Рекомендации по устойчивому лесопользованию на осушаемых землях (44)	3	4
Динамика лесопользования и состояние лесного фонда Карелии (38)	4	7
Механизация восстановления леса в системе интенсивного лесопользования (37)	5	2
Структура лесного фонда, динамика и перспективы лесопользования в Карелии (36)	6	5
Проблемы интенсификации лесопользования в Республике Карелия (36)	7	11
Социальные институты лесного хозяйства, их влияние на эффективность лесопользования (вопросы теории и практики) (33)	8	9
Пространственно-временная динамика лесного фонда и лесопользования европейской части РФ (32)	9	8
Научные разработки Института леса КарНЦ РАН и их реализация в области лесопользования и лесовосстановления (30)	10	13
Системный подход к ключевым проблемам развития экономики лесопромышленного комплекса Республики Карелия (26)	11	14
Выбор технологии лесозаготовок на основе экологической совместимости с лесной средой (23)	12	10
Освоение заболоченных лесов как фактор интенсификации лесопользования в Республике Коми (17)	13	6
Противоречия интеграционных процессов в лесопромышленном комплексе (13)	14	12

Рис. 3 Результаты ранжирования по запросу «68.47.29. Лесопользование. Заготовка древесины»

данный момент не учитывались при вычислении веса публикации).

Результаты проведенного эксперимента (рис.3) были показаны эксперту в данной предметной области, который оценил результаты ранжирования публикаций, полученные нашей системой, как более релевантные его информационным потребностям, по сравнению с результатами Яндекса.

Следует также отметить, что по запросу «68.47.29. Лесопользование. Недревесные лесные ресурсы» поисковая система Яндекс нашла всего 3 документа, а при поиске по таксономии было найдено 17 документов. Такая разница в результатах поиска объясняется тем, что в нашем случае система, при поиске по таксономии, осуществляет автоматическое расширение запроса за счет включения в него терминов таксономии, связанных с ним семантическими отношениями (в данном случае потомками термина «Недревесные лесные ресурсы» являются термины: «Лекарственные», «Пищевые» и т.д.).

4 Заключение

Предложенная онтологически-ориентированная технология систематизации и поиска электронных научных публикаций позволяет на наш взгляд разработать эффективный метод их ранжирования, который учитывает как специфические особенности самой технологии, так и известные приемы построения функций ранжирования, основанные на использовании статистической информации о распределении терминов в запросе и текстах публикаций.

К специфическим особенностям технологии поиска, используемых нами при построении функции ранжирования, относится предположение о том, что **указанные в запросе термины, расположенные на «нижних» уровнях древовидной структуры (представляющей таксономию), в большей степени определяют для пользователя «ценность» публикации, чем термины, расположенные на «верхних» уровнях этого дерева.**

Указанные в запросе термины иерархически связаны между собой определенными отношениями

(классификации, агрегации, часть-целое и синонимии) и в этом смысле они должны оказывать большее влияние на качество ранжирования публикаций, в отличие, скажем, от списка ключевых слов, в котором обоснованное выделение более значимых для целей ранжирования терминов представляется затруднительным делом.

В настоящее время проводится ряд вычислительных экспериментов, результаты которых позволят более точно оценить значения параметров предложенной функции ранжирования (1), а также предложенного метода ранжирования публикаций в целом.

Авторы приносят свои благодарности Ю.В. Чирковой, Н.Б. Луговой и В.Г. Старковой за плодотворное обсуждение рассматриваемых вопросов и реализацию исследовательского прототипа технологии.

Работа выполнена при частичной поддержке гранта РФФИ № 12-07-00070а.

Литература

- [1] Manning, C. An Introduction to Information Retrieval / Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze – Cambridge, England: Cambridge University Press. – April 2009. – P. 544.
- [2] Robertson, S. E., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, Vol. 3, No. 4 (2009) 333–389.
- [3] David Vallet, Miriam Fernández and Pablo Castells An Ontology-Based Information Retrieval Model /Lecture Notes in Computer Science, 2005, Volume 3532/2005, 103–110.
- [4] Raquel Trillo, Laura Po, Sergio Ilarri, Sonia Bergamaschi, Eduardo Mena Using semantic techniques to access web data //Information Systems. 36 (2011). P. 117–133.
- [5] Mauro Dragoni, Cília da Costa Pereira, Andrea G.B. Tettamanzi A conceptual representation of documents and queries for information retrieval system by using light ontologies /Expert Systems with Applications 39 (2012) 10376–10388.
- [6] Добров Б.В., Лукашевич Н.В. Онтология по естественным наукам и технологиям ОЕНТ: структура, состав и современное состояние /Российский научный электронный журнал «Электронные библиотеки», 2008–Том11–Выпуск 1.
- [7] В. Вдовицын, В. Лебедев. Технологии информационного обеспечения научных исследований в ИАС «Природные ресурсы Карелии» // Информационные ресурсы России. № 1. 2012. С. 7–12.
- [8] В.Т. Вдовицын, В.А. Лебедев. Оценка эффективности технологий систематизации и поиска электронной научной информации в ИАС «Природные ресурсы Карелии» // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды 13-й Всероссийской научной конференции RCDL'2011 (Воронеж, 19–22 октября 2011 г.), 2011. С. 309–316.
- [9] В. Вдовицын, В. Лебедев. Технологии систематизации и поиска электронной научной информации с применением онтологий // Информационные ресурсы России. № 5. 2010. С. 6–10.
- [10] Kurt Sandkuhl, Alexander Smirnov, Vladimir Mazalov, Vladimir Vdovitsyn, Vladimir Tarasov, Andrew Krizhanovsky, Feiyu Lin, Evgeny Ivashko Context-Based Retrieval in Digital Libraries: Approach and Technological Framework //Proceedings of the 11th All-Russian Research Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» – RCDL'2009, Petrozavodsk, Russia, 2009. P. 151–157.
- [11] Сайт Рабочей группы Симпозиума «Онтологическое моделирование» URL: <http://ontology.ipi.ac.ru/index..>
- [12] Титов А.Ф., Вдовицын В.Т., Лебедев В.А., Полин А.К. Информационно-аналитическая система поддержки и сопровождения исследований природных ресурсов региона //Труды XII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». RCDL'2010, Казань. 13–16 октября 2010 г. С. 529–534.
- [13] Dr. E. Garcia Tutorial on Okapi Simple BM25F – URL: <http://www.miislita.com/information-retrieval-tutorial/okapi-simple-bm25f-tutorial.pdf> (дата обращения: 18.04.2012).

Document ranking in ontology-based information retrieval system

Vladimir Vdovitsyn, Viktor Lebedev

We propose an algorithm for document ranking produced as part of the work for development of the ontology-based technology for data systematization and retrieval. Both specific characteristics of the technology and some known methods for building ranking functions for retrieval systems based on key words are taken into account.