

Отзыв

на автореферат диссертационной работы Шорина Олега Николаевича

Методы и алгоритмы интеграции большого объема библиографических записей в открытое семантическое пространство»,

представленной на соискание ученой степени кандидата технических наук по специальности 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Диссертация О.Н. Шорина посвящена задачам, возникающим при сборе библиографических записей из разнородных источников, выявлению среди собранного множества записей дубликатов, поиска уже опубликованных данных в открытом семантическом пространстве, с которыми можно установить связь, и установлении связи. Необходимо отметить, что для решения поставленных задач для множеств, содержащих несколько десятков миллионов записей, требуются масштабируемые подходы. Для удовлетворения этого условия автором создается распределенная модель пополнения централизованного хранилища данных, а также описывается алгоритм выявления дублетных записей, который имеет линейную сложность. Автором предложен адаптивный подход разбиения множества записей на кластеры, с помощью которого изменяется критерий принадлежности той или иной записи к кластеру в зависимости от общего количества записей. Также автором доказан ряд теорем, позволяющих оптимизировать процесс подсчета меры близости текстовых полей, основываясь на некоторых легко вычисляемых свойствах этих полей, таких как, например, длина. Разработанная автором система была применена для публикации данных, полученных из электронного каталога Национальной электронной библиотеки. В результате мировое сообщество получило открытый доступ к библиографическим данным библиотек России, пополняемым в автоматическом режиме. Полученные автором результаты являются новыми и важными как с теоретической, так и с практической точки зрения, поскольку они представляют интерес для задач выявления похожих частей информации и могут быть применены, например, в системах выявления плагиата или определения целого произведения по представленному фрагменту.

По тексту автореферата имеются следующие замечания:

1. В автореферате упоминается о создании онтологии предметной области, состоящей из двух классов «Автор» и «Произведение», однако свойства этих классов не представлены.
2. На стр. 9 автореферата говорится, что в диссертации приводится синтаксис запроса

протокола OAI-PMH, посылаемым центральным сервером библиотекам, но как так можно говорить про протокол OAI-PMH, в котором все запросы имеют жестко фиксированный синтаксис.

3. Код ISBN состоит не только из цифр, но в нем может присутствовать буква X, и как тогда работает алгоритм сравнения.
4. Наиболее частая ошибка в каталоге – это ошибка в поле «год издания».

Отмеченные недостатки не умаляют достоинств работы и относятся к стилистическим недочетам. Автореферат написан понятным языком, хорошо прослеживается логика излагаемого материала. Результаты работы опубликованы в рецензируемых журналах и апробированы на авторитетных конференциях и семинарах. Судя по автореферату, диссертация соответствует специальности 05.13.11 и удовлетворяет требованиям, предъявляемым к кандидатским диссертациям. Считаю, что Шорин О.Н. достоин присуждения ему ученой степени кандидата технических наук.

Член-корреспондент РАН
Федотов Анатолий Михайлович
д.ф.-м.н., профессор,
главный научный сотрудник

Федеральное государственное бюджетное учреждение науки Институт вычислительных технологий
Сибирского отделения Российской академии наук,
630090, Новосибирск, пр. ак. Лаврентева,6,
тел. +7 (383) 330 7351
e-mail: fedotov@sbras.ru

21.02.2017

Подпись Федотова А.М. заверяю
Врио директора ИВТ СО РАН, к.ф.-м.н.

