

МЕТОД СРАВНЕНИЯ RDF ТРИПЛЕТОВ

Куриной И.В., Штогрин Е.С.

Институт телекоммуникационных систем НТУУ «КПИ», Украина

E-mail: kur91@mail.ru

Аннотация — В работе предложен метод сравнения RDF триплетов. Метод позволяет определить являются ли дубликатами не только полностью идентичные триплеты, а и те в которых были допущены синтаксические ошибки, или использован разный формат записи названий составляющих частей триплета.

1. Введение

Одним из наиболее распространенных способов представления семантических данных является *Resource Description Framework (RDF)*. Это способ описания ресурсов с помощью триплетов «субъект — предикат — объект» [1]. В *RDF* возможно появление дубликатов из-за неточного ввода информации или при получении информации из различных источников. В [2] предложен метод поиска дубликатов, однако он не учитывает возможность поиска не полностью совпадающих триплетов. В данной работе рассматривается метод, позволяющий проводить сравнение *RDF* триплетов, в которых присутствуют синтаксические ошибки или используются различные форматы записи.

2. Основная часть

Для поиска *RDF* дубликатов необходимо определить метод сравнения двух триплетов. В работе предлагается проводить сравнение следующим образом: для каждой пары *RDF* триплетов отдельно сравниваются субъект, предикат и объект соответственно. Результатом сравнения является степень совпадения субъекта, предиката и объекта.

RDF триплет характеризуется логической завершенностью, цельностью, отдельные его части — субъект, предикат и объект, представляют информацию, только будучи связанными в триплет, поэтому требуется вычислить показатель, который будет определять степень совпадения всего *RDF* триплета в целом.

Далее требуется сравнить значение полученного показателя со значением, которое показывает, следует считать *RDF* триплеты дубликатами или нет.

Появление дубликатов *RDF* триплетов может быть вызвано тем, что триплеты записаны на разных языках или с использованием транслитерации, однако представляют собой одну и ту же информацию. А также в случае, когда язык триплетов совпадает, но при записи триплетов были допущены синтаксические ошибки, или использован разный формат записи.

В случае, когда для записи названий субъекта, предиката и объекта триплетов используются разные языки, для сравнения триплетов следует определить языки, на которых они записаны, а затем использовать методы сравнения с использованием словаря.

В случае записи триплетов с использованием одинакового языка для сравнения названий субъекта предиката и объекта можно воспользоваться методами сравнения текстовых строк и расширить их с учетом особенностей *RDF* триплетов, которые описаны выше.

Для сравнения текстовых строк существуют такие методы: расстояние Левенштейна, алгоритм Смита–Ватермана, метод N-грамм, алгоритм Soundex [3].

В данной работе за основу взят алгоритм Смита–Ватермана, который определяет количество совпа-

дающих символов в двух строках. Данный метод выбран потому что он более эффективен при сравнении слов разной длины, а также имеет наименьшую вычислительную сложность и затраты ресурсов памяти.

Предлагаемый метод состоит из нескольких этапов описанных ниже. На первом этапе определяется количество совпавших символов K каждой пары субъектов, объектов и предикатов. На следующем шаге вычисляется отношение соответствия $S = \frac{K}{N}$

количества совпавших символов K к длине N более длинного слова для субъекта, предиката и объекта триплета соответственно. В результате эти отношения соответствия перемножаются, и получается показатель сходства для двух триплетов $S_{тр} = S_{суб} \cdot S_{пред} \cdot S_{об}$. Полученный показатель сравнивается с уровнем, который показывает, можно ли считать *RDF* триплеты дубликатами или нет.

В случае, когда триплеты написаны с использованием транслитерации, нужно привести записи к одному языку, используя карту или таблицу соответствия транслитированных символов. Дальнейшие действия сравнения *RDF* триплетов проводятся согласно предложенному методу для случая совпадения языков триплетов.

3. Заключение

В данной работе предложен метод сравнения *RDF* триплетов, позволяющий выдавать предположения о равенстве двух триплетов. Применение данного метода позволит находить дубликаты *RDF* триплетов, тем самым улучшая качество информации, получаемой из различных источников.

Дальнейшая работа будет направлена на определение и обоснование значения уровня, с которым сравнивается показатель, определяющий являются ли триплеты дубликатами. А также на разработку метода поиска дубликатов во множестве *RDF* триплетов с использованием минимального количества операций сравнений *RDF* триплетов.

4. Список литературы

- [1] RDF / The World Wide Web Consortium. — <http://www.w3.org/RDF/>. — 04.02.2013.
- [2] Zakia S. Detection and elimination of duplicate data from semantic web queries / S. Zakia // International Journal of Multidisciplinary sciences and Engineering. — 2010. — Vol.1, № 2. — P. 1 — 6.
- [3] Elmagarmid A.K. Duplicate record detection: A Survey / A.K. Elmagarmid, P.G. Ipeirotis, V.S. Verykios // IEEE Transactions on Knowledge and Data engineering. — 2007. — Vol. 19, №1. — P.1 — 16.

RDF TRIPLES COMPARISON METHOD

Kurinnoy I.V., Shtogrina O.S.

Institute of Telecommunication Systems

National Technical University of Ukraine "KPI", Ukraine

Abstract — The method of the RDF triples comparison and the duplicates detection is proposed. The proposed method make it possible to detect duplicates not only for totally similar ones, but when some of them have mistakes or have different formats.