8. Любченко, В.В. Метод поиска фрейма по шаблону на основе ассоциаций [Текст]/ В.В. Любченко, В.А. Крисилов // Труды Одесского политехнического университета, спецвыпуск. – Одесса, Издательство Одесского политехнического университета, 2006. – С. 60 – 63.

Д.В. Еленев, А.О. Линник

АВТОМАТИЗИРОВАННОЕ СНИЖЕНИЕ ДУБЛИРОВАНИЯ ИНФОРМАЦИИ В СИСТЕМЕ МОНИТОРИНГА ДЕЯТЕЛЬНОСТИ ПОДРАЗДЕЛЕНИЙ УНИВЕРСИТЕТА

(Самарский государственный аэрокосмический университет имени академика С.П.Королева (национальный исследовательский университет))

Задача сбора и анализа данных о результативности деятельности подразделений является насущной для всех крупных научно-образовательных организаций. Построение внутренних рейтингов подразделений и работников, оценки результатов научной деятельности работников и студентов, подготовка внешних отчетов требует полноту, достоверность и своевременность поступления данных.

В Самарском государственном аэрокосмическом университете имени академика С.П.Королева (национальном исследовательском университете) задача сбора сведений о достижениях работников и подразделений решается при помощи информационно-аналитической системы мониторинга деятельности подразделений и количественной оценки качества работы университета [1]. В состав системы мониторинга входит комплекс автоматизированных рабочих мест (АРМ): «Ректор», «Проректор», «Дирекция Программы развития национального исследовательского университета», «Кафедра», «Управление образовательных программ», «Интеллектуальная собственность», «Научно-исследовательская работа студентов», «Управление обеспечения инновационной деятельности», «Отдел управления качеством» и др. Основная нагрузка по сбору сведений о достижениях работников и подразделений ложится на АРМ «Кафедра».

Для оперативного контроля деятельности подразделений в системе мониторинга используется построение отчетов на основании данных, введенных в течение текущих контрольных периодов. Текущий контрольный период установлен равным декаде, а текущие контрольные точки – 10, 20 и последние числа каждого месяца. Данные, собранные и обобщенные пользователем АРМ за текущий контрольный период, заносятся пользователями АРМ «Кафедра» не позднее даты текущей контрольной точки.

В число предоставляемых пользователями АРМ «Кафедра» данных входят сведения об опубликованных научных работах, участии студентов в конкурсах, участии работников в конкурсах, проведенных научных мероприятиях,

результатах анкетирования работников и студентов, участии в выставках и подготовке выставочных экспонатов, материальной базе и т.д.

Последующая за вводом обработка данных, расчет интегральных показателей и построение аналитических отчетов осуществляются в системе автоматически. Для построения отчетов используются те данные, достоверность которых была подтверждена, во-первых, пользователем АРМ «Кафедра» путем закрытия отчетного периода и, во-вторых, пользователем соответствующей точки ответственности в системе (согласно возложенной ответственности по учету данных). Например, для сведений об опубликованных научных работах точкой ответственности является отдел сопровождения научных исследований с одноименным АРМ. По результатам проверки введенные сведения могут быть подтверждены, отклонены или отмечены как дублирующиеся.

Достоверность и чистота данных в системе мониторинга являются важным фактором, позволяющим их использование в построении рейтингов работников и подразделений и для вычисления баллов (по ряду показателей) профессорско-преподавательского состава в системе стимулирования труда ППС, действующей в СГАУ с 2013 г.

Основными причинами дублирования информации являются повторный ввод сведений одним и тем же пользователем и предоставление одних и тех же сведений несколькими пользователями (работниками разных кафедр). Наиболее характерным с точки зрения дублирования видом данных являются сведения об опубликованных научных работах, а основным признаком дублирования – схожесть названий публикаций. В случае предоставления сведений об одной и той же публикации полного совпадения названий зачастую не происходит из-за опечаток, копирования знаков переноса и т.п., что осложняет поиск и учет дублированных записей.

Ранее в системе мониторинга для поиска в существующей базе данных был разработан специализированный алгоритм, использующий коэффициент сходства предполагаемых дубликатов записей с подобными записями [3], работающий на основе алгоритма Metaphone. Сравнение не самих названий, а сгенерированных на их основе ключей позволяет уменьшить влияние опечаток на результаты поиска, т.к. схожие по звучанию слова дают одинаковые ключи. В то же время данный разработанный алгоритм слабо справляется с поиском записей, имеющих пропущенные слова и не позволяет избавляться от дублирования на стадии ввода информации.

Автоматизация снижения степени дублирования может быть разделена на две взаимодополняющих части: усовершенствованный поиск дубликатов среди уже имеющихся записей и поиск записей при вводе новых записей в базу данных системы.

Для сравнения строк были опробованы три алгоритма: алгоритм шинглов (англ. shingles – чешуйки) и алгоритмы на основе метрик Дамерау-Левенштейна и Джаро-Винклера. Перед началом работы каждого алгоритма строки приводятся к общему виду: символы переводятся в нижний регистр, убираются служебные символы, кратные пробелы, пробелы в начале и конце строк.

В алгоритме шинглов обе строки разбиваются на «чешуйки» длиной 2 слова каждая (если хотя бы одна строка состоит из одного слова, то длина «чешуек» становится равной 1). «Чешуйки» составляются внахлест. Для каждой «чешуйки» высчитывается циклическая контрольная сумма и производится сравнение полученных контрольных сумм. Схожесть S двух строк находится здесь как

$$S = \frac{2a}{n+m},$$

где α – количество совпавших контрольных сумм, n – количество слов в первой строке, m – количество слов во второй строке. Алгоритм не учитывает возможные ошибки в словах, и из-за того, что сравнивает наборы слов, не способен полноценно распознать схожесть строк, в которых слова будут переставлены местами.

В алгоритме на основе метрики Дамерау-Левенштейна используется так называемое расстояние Дамерау-Левенштейна – количество символов, которые нужно удалить, вставить или заменить в одном слове, чтобы получить из него другое слово. Две строки, поданные на вход программы, разбиваются на два массива по словам, в качестве разделителя используется пробел. После чего для каждого слова первой строки вычисляется расстояние Дамерау-Левенштейна ко всем словам второй строки, тем самым строя матрицу размером (n x m), где n и m – количество слов первой и второй строки соответственно.

В алгоритме на основе метрики Джаро-Винклера используется расстояние Джаро-Винклера D_w между двумя словами, которое находится по формуле:

$$D_j = \left(\frac{m}{a} + \frac{m}{b} + \frac{m-t}{m}\right) * \frac{1}{3}, D_w = \begin{cases} D_j, & \text{если } D_j < 0{,}7, \\ D_j + \left(0{,}1 * l\left(1 - D_j\right)\right), & D_j \geq 0{,}7, \end{cases}$$

где D_j – расстояние Джаро, a – длина первого слова, b – длина второго слова, m – количество подходящих знаков, t – количество перестановок, D_w – расстояние Джаро-Винклера, l – количество первых совпадающих символов в обоих словах. Количество перестановок – это количество подходящих символов, находящихся в неправильном порядке, деленное на два.

Подходящими знаками слов считаются знаки, если они равны и находятся друг от друга не далее, чем

$$\left[\frac{\max(a,b)}{2}\right] - 1.$$

Аналогично алгоритму на основе метрики Дамерау-Левенштейна строится матрица, хранящая в себе расстояния Джаро-Винклера. Два слова совпадают полностью, если расстояние Джаро-Винклера для них равно единице.

Схожесть S двух строк в алгоритмах на основе метрик Дамерау-Левенштейна и Джаро-Винклера высчитывается как

$$S = \frac{4n+m}{2k},$$

где n - суммарное количество букв в совпавших словах, m - суммарное количество букв в частично совпавших словах, k - суммарное количество букв в словах обеих строк.

На рис. 1 приведена тестовая форма поиска сходных записей на основе метрики Дамерау-Левенштейна. Анализ результатов работы вышеописанных алгоритмов показал хорошее совпадение его результатов с результатами проверки записей об опубликованных научных работах пользователем соответствующего АРМ.

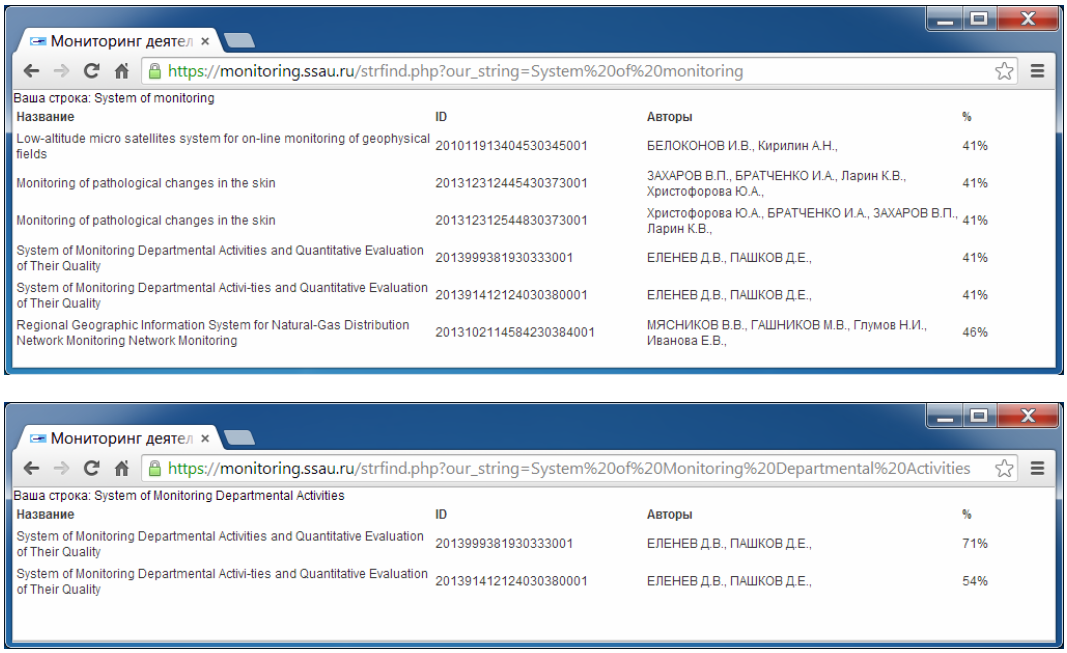

Рис. 1. Результаты поиска сходных записей

Необходимо отметить, что, если в сравниваемых строках присутствуют частично совпадающие слова, то алгоритмы на основе метрик Дамерау-Левенштейна и Джаро-Винклера на реальных примерах зачастую показывают низкую степень схожести при высокой смысловой. Например, при сравнении строк «Определение длины световой волны» и «Определения длин световых волн» оба алгоритма возвращают схожести всего 50%, потому что все слова в этих двух строках совпадают лишь частично. Для дальнейшего повышения эффективности поиска сходных записей возможно проведение дополнительной канонизации сравниваемых строк (преобразование символов в сходные по звучанию, удаление предлогов) с последующим применением алгоритмы Metaphone.

Литература

1. Еленев Д.В., Кузьмичев В.С., Пашков Д.Е. Автоматизация системы управления национальным исследовательским университетом и мониторинга его деятельности // Программные продукты и системы. - 2012. - № 3. - С. 31-34.

2. Информационная инфраструктура инновационного вуза. Опыт СГАУ: монография / [А.В.Баскаков и др.] – Самара: Изд-во СамНЦ РАН, 2013. – 124 с.

3. Еленев Д.В., Ризванова Л.Н. Автоматизация поиска подобных записей в базе данных опубликованных научных работ // Международный научно-технический форум, посвященный 100-летию ОАО «Кузнецов» и 70-летию СГАУ, Самара, 5-7 сентября 2012 года: сборник трудов в 3-х томах. Том 3. Всероссийская молодежная научно-техническая конференция «Космос-2012». –

Самара: Издательство Самарского государственного аэрокосмического университета, 2012. - С. 169-171.

А.В. Иващенко, И.А. Сюсин

АНАЛИЗ РИТМИЧНОСТИ НАЗНАЧЕНИЯ ПРИ УПРАВЛЕНИИ ПОСРЕДНИЧЕСКОЙ ДЕЯТЕЛЬНОСТЬЮ В СФЕРЕ УСЛУГ

(Самарский государственный аэрокосмический университет имени академика С.П. Королева (национальный исследовательский университет))

Организация виртуального посреднического оператора для управления взаимодействием поставщиков и потребителей различных услуг в едином информационном пространстве позволяет задать правила такого взаимодействия, не лишая при этом отдельных его участников самостоятельности в принятии решений. В этой ситуации поставщики и потребители услуг образуют виртуальное сообщество, обладающее свойствами самоорганизации [1], которое представляет собой интересный объект для исследования. Управление посреднической деятельностью в таком сообществе может быть автоматизировано с использованием современных подходов по ин-формационному управлению в интегрированной информационной среде [2, 3].

Поставщики и потребители услуг в процессе достижения своих целей постоянно выстраивают новые отношения и вступают в информационное взаимодействие. В частности, некоторые из них могут выступать в качестве посредников, получая определенную выгоду от выстраивания отношений между членами виртуального сообщества путем передачи соответствующей информации. Однако обратной стороной свободы действий участников взаимодействия является высокая степень неопределенности их действий – трудно предсказать, когда поток заявок на получение услуги будет выше или ниже среднего прогнозируемого значения.

Такой характер процесса получения заказов является причиной определенных трудностей по автоматизации процессов работы поставщика услуг, планирования его деятельности (особенно долгосрочного), выстраивание отношений с другими поставщиками. Возникает необходимость вносить постоянные изменения в планы, а одновременное проведение нескольких изменений вызывает сложность в синхронизации и необходимость борьбы с тупиковыми ситуациями [4].

Для решения этих проблем, а также с целью повышения эффективности использования оптимизационных алгоритмов планирования, предлагается в состав автоматизированной системы реализации виртуального посреднического оператора включить модуль по предварительной обработке входных очередей событий на каждом этапе планирования заказов на получение услуг. Это позволит синхронизировать деятельность различных поставщиков и обеспечить их согласованное взаимодействия в условиях неравномерной динамики внешних событий.