программные агенты, как раз и определяет адекватность такого выбора.

Библиографический список

1. Кудрявцев Д. Технологии применения онтологий [Электронный ресурс] //Бизнес Инжиниринг Групп: сайт. – Режим доступа: http://bigc.ru/theory/km/onto_technologies.php.

2. Добров Б.В. Онтологии и тезаурусы [Электронный ресурс] // Основан на курсе Intuit.ru «Онтологии и тезаурусы: модели, инструменты, приложения». Авторы: Б.В. Добров, В.В. Иванов, Н.В. Лукашевич, В.Д. Соловьев. – Режим доступа: http://www.intuit.ru/studies/courses/1078/270/info.

3. Морозов А.А. Об одном подходе к логическому программированию интеллектуальных агентов для поиска и распознавания информации в Интернет [Электронный ресурс] // Журнал радиоэлектроники. – № 10, 2003. – Режим доступа: http://jre.cplire.ru/iso/nov03/1/text.html

4. Морозов А. А., Обухов Ю. В. Акторный Пролог [Электронный ресурс] //электронная книга (версия от 23.01 2004). – Режим доступа: http://www.cplire.ru/Lab144/aprolog.pdf, свободный.

УДК 004.896

Алгоритм семантического поиска в больших текстовых коллекциях

В.В. Савченко, Е.Н. Крючкова
АлтГТУ, г. Барнаул

Проблема поиска в больших текстовых коллекциях является одной из приоритетных в условиях большого и стремительно растущего объема информации [1]. Одним из вариантов поиска является семантический поиск, т.е. поиск по смыслу содержащейся в тексте информации [2–4]. Существующие системы поиска (Google, SearchMonkey, Powerset, Freebase, AskNet) имеют существенные ограничения на длину запроса, демонстрируют снижение качества поиска с увеличением поискового запроса, имеют лишь незначительное улучшение результатов поиска при использовании семантики. Кроме того, большинство таких поисковых систем работают только с английским языком.

Изменчивость синтаксических конструкций и вариативность лексики естественных языков, разнообразие стилей изложения материала существенно усложняют решение данной задачи.

Семантический поиск. В данной работе представлен результат разработки системы семантического поиска для больших текстовых коллекций на русском языке. Ключевая особенность предлагаемой системы – снятие ограничений на величину поискового запроса.

Исходными данными для поиска являются текстовые коллекции и запрос пользователя, который также представляет собой текстовую коллекцию. Логично сделать вывод, что большая текстовая коллекция в общем случае неоднородна по своему содержанию и при поиске интересна лишь ее определенная часть, текст можно разделить на фрагменты. Как правило, такие фрагменты – это страницы, абзацы или наборы из нескольких предложений. Фрагменты будем называть «окнами», таким образом, задача сводится к поиску таких окон.

Для каждого окна запроса и поисковых коллекций строится граф семантических связей – «семантический граф». Семантический граф представляет собой ориентированный граф, вершинами которого являются слова русского языка, представленные в нормальной форме, а ребра характеризуются весом и типом семантической связи. Направление ребра зависит от типа семантической связи, например, дуга может соответствовать отношению «объект – действие», «объект – свойство», «действие – время». Для построения семантического графа каждое предложение из окна коллекции обрабатывается семантическим анализатором. В данной работе используется семантический анализатор RML (http://www.aot.ru).

Предложения окна обрабатываются последовательно, для каждого очередного предложения строится семантический граф G_{new}. Все ребра семантического графа G_{new} первоначально имеют вес равный единице. На каждой итерации семантический граф G_i предыдущей итерации объединяется с графом G_{new} обрабатываемого предложения, причем веса ребер нового графа G_{i+1} вычисляются как сумма весов ребер графов G_i и G_{new}, что приводит к увеличению весов однотипных ребер. После этого результирующий семантический граф G_{i+1} используется для следующей итерации.

После обработки всех предложений окна, веса ребер получившегося графа нормируются путем деления на общее количество семантических связей окна.

Для расчета коэффициента релевантности окна запроса и окна коллекции необходимо определить соответствие ребер графа окна и графа коллекции. Однако один и тот же смысл содержится в текстах разного

стилевого оформления, вершины графа запроса и графа окна коллекции могут быть не только синонимами, но и каким-либо образом связанными по смыслу словами, могут содержать обобщающие сведения или только частичную информацию. Например, при обработке запроса, связанного с поиском текста «*хищные животные, обитающие в тайге*» к хищникам, обитающим в тайге, в том числе относятся и лисы. Однако между хищниками и лисами не должно быть полного отождествления, т.к. хищники – это не только лисы. Отсюда возникает необходимость учета семантической зависимости при поиске соответствующих ребер графов.

Для поиска связанных по смыслу слов был использован словарь, полученный методом автоматической обработки краткого толкового словаря, в котором представлен перечень слов в нормальной форме [5]. Каждому слову сопоставлен набор слов, связанных с ним ассоциативной, синонимичной и т.д. связью. Таким образом, словарь представляет собой направленный граф $G_{word} = (V_{word}, U_{word})$, где вершины V_{word} - это слова в нормальной форме, а ребра U_{word} имеют действительные весовые коэффициенты от 0 до 1. Назовем граф G_{word} графом справочника. За коэффициент связанности слов a_k и a_m - вершин семантического графа запроса $G_{request}$ и семантического графа окна коллекции G_{text} возьмем произведение весов от таких же слов до общего предка в графе G_{word}. При совпадении слов данный коэффициент будет равен 1, иначе будет принадлежать промежутку [0;1], т.к. веса ребер графа G_{word} - действительные числа от 0 до 1.

Граф G_{word} является упрощенной моделью наших знаний о реальном мире, а общий предок слов a_k и a_m в этом графе представляет собой некоторое обобщение соответствующих понятий.

За коэффициент соответствия ребра графа $G_{request}$ и ребра графа G_{text} возьмем произведение коэффициентов связности соответствующих вершин, при условии, что типы семантических связей ребер совпали. Иначе коэффициент соответствия ребер равен 0.

Значение коэффициента релевантности определяем по формуле 1:

$$S = \sum_{k=1}^{n} \underset{t=1}{\overset{m}{MAX}} (C_t^k l_k j_k), \qquad (1)$$

где n – количество ребер графа $G_{request}$, m – количество ребер графа G_{text}, C_t^k – коэффициент соответствия ребера k графа $G_{request}$ и ребра t графа G_{text}, l_k – вес ребра k семантического графа $G_{request}$, j_t – вес ребра t семантического графа G_{text}. Определив максимальное значение по всем окнам текстовой коллекции, получим общее значение – коэффициент релевантности запроса и текстовой коллекции.

Тесты и результаты. Для оценки полученной системы были отобраны текстовые коллекции большого объема, удовлетворяющие одному и тому же запросу к поисковой системе google.ru. Поисковый запрос строился исходя из содержания определенных коллекций. Для коллекций, по содержанию которых строился запрос или коллекций аналогичного содержания, значение коэффициента превосходило значение коэффициента других коллекции, отличающихся по содержанию.

Временные затраты на обработку коллекции в зависимости от количества слов в запросе и окне коллекции представлены на рисунке 1.

Рис. 1. Временные затраты

Так же было проведено тестирование с добавлением «шума», т.е. добавлением в текст окна предложений из других текстовых коллекций, не соответствующих запросу. Результаты отклонения коэффициента релевантности приведены на рисунке 2.

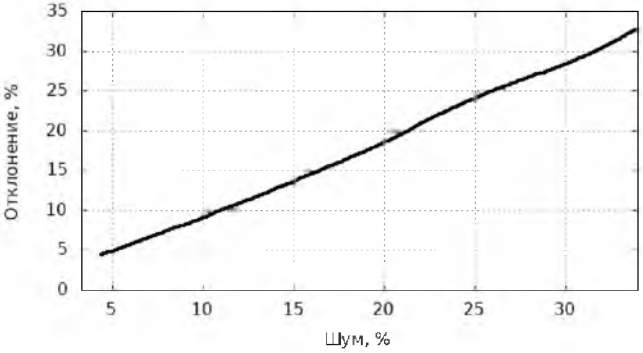

Рис. 2. Отклонение коэффициента релевантности из-за шума

Эксперименты показали, что отклонение коэффициента релевантности линейно возрастает с увеличением шума, т.е. алгоритм расчета коэффициента корректен и действительно отображает соответствие между окном коллекции и поисковым запросом.

Параллельная обработка окон текстовой коллекции на многопоточных и многопроцессорных системах выполняется параллельно, что значительно повышает скорость обработки запроса.

Библиографический список

1. Hannah Bast, Marjan Celikik Efficient Fuzzy Search in Large Text Collections // ACM Transactions on Information Systems, 2010.

2. Mathieu d'Aquin, Enrico Motta Watson, more than a Semantic Web search engine // IOS Press Amsterdam, 2011.

3. K. Elbedweihy, S.N. Wrigley, F. Ciravegna, D. Reinhard, A. Bernstein Evaluating Semantic Search Systems to Identify Future Directions of Research // Second International Workshop on Evaluation of Semantic Technologies, 843, page 25-36, 2012.

4. G. Tsoumakas, M. Laliotis, N. Markantonatos, I. Vlahavas Large-Scale Semantic Indexing of Biomedical Publications at BioASQ // BioASQ Workshop, 2013.

5. Крайванова В.А., Кротова А.О., Крючкова Е.Н. Построение взвешенного лексикона на основе лингвистических словарей // ЗОНТ-2011 : материалы Всероссийской конференции с международным участием. Т. 2. – Новосибирск, 2011. – С. 32–38

УДК 528.236

Пересчет объектов капитального строительства из региональных систем координат в систему координат World Geodetic System

С.И. Суханов
АлтГУ, г. Барнаул

В качестве объектов капитального строительства могут выступать волоконно-оптические линии связи, линии электропередач, электроподстанции, и т.д. У многих крупных предприятий энергетики данные объекты были поставлены на государственный кадастровый учет в нескольких регионах одновременно. В разных регионах России для ведения государственного кадастрового учета были приняты свои сис-