Использование семантического анализа текстов для поиска специалистов

Игорь Захлебин

НИУ ВШЭ, Москва, Россия
zahl.igor@gmail.com

Аннотация В работе предложен метод семантического поиска специалистов по набору составленных ими текстов. Описан формат запросов, позволяющий определять набор искомых компетенций. Разработаны алгоритмы построения и сравнения семантических представлений фрагментов текстов на естественном языке. На основе предложенной модели разработан и испытан прототип поисковой системы ExpSearch-1 (Experts Search, версия 1).

Ключевые слова: поиск специалистов, семантический анализ, теория К-представлений, естественно-языковые запросы.

1 Введение

Перед современными компаниями остро стоит проблема поиска квалифицированных специалистов. При этом поиск приходится осуществлять не только среди кандидатов на открывшиеся позиции, но и среди собственных сотрудников, например, для устранения нештатных ситуаций [1]. Поэтому для повышения эффективности бизнеса повсеместно разрабатываются автоматические системы, позволяющие ускорить и качественно улучшить этот процесс. Так, компания IBM менее чем за 6 лет сэкономила около $500 миллионов благодаря внедрению собственной системы поиска персонала [2].

Среди методов, применяющихся для поиска специалистов, наиболее популярным остается поиск по ключевым словам. Как правило, менеджер по персоналу, имеющий базу резюме специалистов, с помощью специального программного обеспечения осуществляет поиск по названию профессии, названиям технических средств и/или профилю образования специалиста. При этом, даже если запрос составлен удовлетворительно:

- при поиске не будут учтены смысловые отношения между словами;
- поисковая выдача будет различаться для запросов с одинаковым значением, но составленных по-разному (даже при учете синонимии понятий);
- оказывается невозможным алгоритмически определить, является ли конкретный пункт выдачи действительно релевантным запросу.

Для устранения перечисленных недостатков в данной работе предлагается модель системы поиска, использующая семантический анализ текстов и оценку релевантности результатов, основанную на сравнении семантических представлений фрагментов текстов.

2 Структура разработанной системы

Для реализации нового подхода к семантическому поиску специалистов была разработана система-прототип ExpSearch-1 (Experts Search, версия 1). Ее структура изображена на рис. 1.

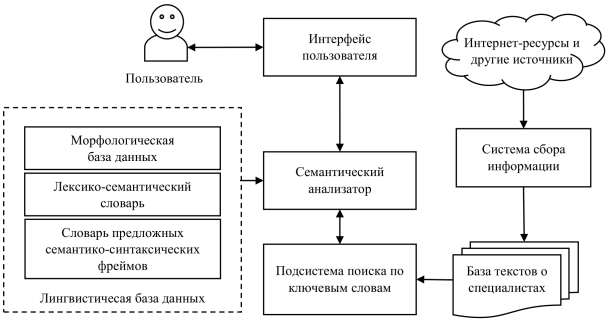

Рис. 1. Схема разработанной системы семантического поиска

В систему загружаются полные тексты с информацией о специалистах (анкеты, резюме, профессиональная переписка и т.п.), которые группируются по принадлежности к соответствующим специалистам.

Для поиска пользователь вводит запрос в виде набора словосочетаний, каждое из которых определяет одну искомую компетенцию [2]. Запрос состоит из существительных с возможным использованием предлогов, прилагательных и числительных. Запрос такого вида позволяет задать, например:

- область знания (эпизодическая логика, управление рисками);
- модель, теорию, понятие (модель Эрроу-Дебре, дефлятор ВВП);
- инструментальное средство (среда SPSS, пакет MatLab);
- умение или навык (обработка древесины, разработка под iOS).

Система ищет специалистов, у которых в связанных с ними текстах присутствуют релевантные словосочетания. Чем большему числу критериев удовлетворяет специалист, тем выше он располагается в ранжировании, выдаваемом системой.

3 Алгоритм построения семантических представлений

В основе работы системы лежит модифицированный алгоритм построения семантических представлений (далее – СП) и модель лингвистической базы данных, предложенные в книгах [3,4].

Построение СП фрагмента текста начинается с определения морфологических свойств его слов и приведения их к начальной форме. Затем к существительным применяется лексико-семантический словарь. По начальной

форме слова он сопоставляет ему семантическое значение sem (для равнозначных с точки зрения системы слов оно совпадает) и набор характеристик $st_1, \ldots, st_k$. Словарь содержит записи вида $(lec, sem, st_1, \ldots, st_k)$, где lec – базовая форма слова; sem – строка, обозначающая семантическое значение лексемы lec; $st_1, \ldots, st_k$ – различные семантические характеристики сущности, связанные с понятием sem; k – наибольшее возможное их число.

Далее к существительным применяется словарь предложных семантико-синтаксических фреймов, задающий связи между семантическими единицами, выделенными на предыдущем этапе. Он содержит записи вида $(prep, st_1, st_2, grc, rel)$, где $prep$ – необходимый предлог (может быть пустым); st_1, st_2 – семантические характеристики, которые можно связать с первым и вторым существительным в лингвистически правильном словосочетании «сущ.1 + $prep$ + сущ.2» соответственно; grc (grammatic case) – обозначение падежа второго существительного; rel – обозначение смыслового отношения. Существительные попарно проверяются на соответствие следующим условиям: первому существительному сопоставлен сорт sr_1, второму – sr_2, зависимое существительное находится в падеже grc, и между ними есть предлог $prep$. При удовлетворении всех условий для одной записи словаря считается, что между существительными установлено смысловое отношение rel из этой записи, дальнейшая сверка по словарю для этой пары прекращается.

Заметим, что на предыдущих шагах обрабатывались только существительные. Если имеется прилагательное или слово, ведущее себя как прилагательное, рассматриваются существительные, между которыми оно расположено. При совпадении с одним из них по роду, числу и падежу оно обозначается зависимым от него. При совпадении с обоими существительными прилагательное считается зависимым от последнего из них по порядку. В данных случаях устанавливается отношение rel – «свойство», а значение sem зависимой единицы – как начальная форма зависимого слова.

В результате выполнения алгоритма получается СП фрагмента текста – ориентированное дерево, в вершинах которого находятся семантические единицы sem, а ребра заданы отношениями rel.

4 Алгоритм поиска

Задачей алгоритма поиска является нахождение фрагментов в текстах о специалистах, имеющих СП (семантические представления), схожие с СП поискового запроса. Поэтому при поиске сначала строится СП запроса, а затем для каждого введенного пользователем словосочетания система составляет набор слов, парные вхождения которых в текст могут потенциально содержать между собой отношения как в СП запроса. Для этого в список включаются все слова из лексико-семантического словаря, которым могут быть сопоставлены единицы sem из СП запроса. Например, для словосочетания «маркетинг сбыта» может быть составлен следующий набор ключевых слов: «маркетинг», «маркетолог», «анализ рынка», «исследование рынка», «сбыт», «продажа».

По текстам, содержащимся в базе знаний, производится поиск по составленному набору ключевых слов. При нахождении хотя бы одного слова строится СП фрагмента текста вокруг него (границы определяются по таким символам, как точка, точка с запятой, табуляция, перенос строки и т.п.). Полученные представления группируются по специалистам, к текстам которых они относятся. Такой подход позволяет сохранить общую вычислительную сложность алгоритма низкой, так как процедура построения СП запускается только на предположительно релевантных фрагментах текстов.

Каждое СП можно упрощенно представить в виде набора триплетов вида (sem_1, rel, sem_2), то есть пар связанных семантических значений. Пусть $\mathbf{A}$ – набор триплетов, представляющих поисковый запрос, а $\mathbf{B}$ – аналогичный набор, представляющий СП, выделенные в текстах, связанных с одним специалистом. Тогда мерой релевантности специалиста будет величина $score = \frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A}|}$, находящаяся в отрезке $[0, 1]$. Для получения результирующего ранжирования специалисты упорядочиваются по убыванию показателя $score$, и их список возвращается пользователю как результат поиска.

5 Заключение

На основе предложенной модели на языке программирования Python был разработан прототип поисковой системы ExpSearch-1. В качестве тестовых данных в систему была загружена текстовая информация о более чем 7000 сотрудниках Высшей школы экономики (НИУ ВШЭ), взятая с официального сайта. В ходе испытаний система успешно выполнила поиск по набору тестовых запросов и дала по ним релевантные результаты.

В качестве направлений для продолжения работы предполагается усложнение формата поддерживаемых запросов и совершенствование алгоритма сравнения семантических представлений.

Список литературы

1. *Xiaodan Song и др.* ExpertiseNet: Relational and Evolutionary Expert Modeling. // SmallBlue Internet Edition (alpha) [Электронный ресурс]. URL: `http://smallblue.research.ibm.com/publications/ExpertiseNet_UM.pdf` (дата обращения: 10.03.2014).
2. *Arjen P. de Vries.* Expert Finding = Finding People + Assessing Expertise // Future Challenges in Expertise Retrieval, SIGIR 2008 Workshop, Singapore [Электронный ресурс]. URL: `https://app.box.com/s/9yqrk9zs6lc38gqpsi2j` (дата обращения: 10.03.2014).
3. *Фомичев В.А.* Формализация проектирования лингвистических процессоров – М.: МАКС Пресс, 2005.
4. *Fomichov V.A.* Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms. Series: IFSR International Series on Systems Science and Engineering, Vol. 27. Springer: New York, Dordrecht, Heidelberg, London, 2010.

Searching for Experts Using the Semantic Analysis of Texts

Igor Zahlebin

Higher School of Economics, Moscow, Russia
zahl.igor@gmail.com

Abstract. This paper presents a semantic method for searching for the experts. The method operates over a set of texts authored by themselves. The query format allowing one to define a set of the selected skills, and the algorithms for constructing and comparing the semantic representations are also presented. The ExpSearch-1 (Experts Search, version 1) system which is based on the present method has been developed and evaluated.

Keywords: experts' search, semantic analysis, K-representations, natural language queries.