Resumagic: система автоматической обработки резюме

© А.В. Сафронов

ООО «Хэдхантер»
safronov@hh.ru

Аннотация

В статье описывается система Resumagic, предназначенная для автоматической обработки резюме. Рассмотрены задачи фильтрации и классификации резюме. Приведены результаты применяемых методов.

1 Введение

В своей повседневной работе специалисты по подбору персонала часто сталкиваются с задачей обработки потока входящих резюме. Информация о соискателях попадает к рекрутеру из нескольких разных источников, в том числе и по электронной почте. В больших кадровых агентствах количество резюме, присылаемых соискателями в течение дня, может достигать нескольких сотен; к этому числу на практике следует также прибавить некоторое количество незапрашиваемой корреспонденции (спама). Каждое входящее письмо нуждается в трудоемкой обработке.

Для автоматизации процесса обработки резюме была разработана система Resumagic. Ее основные функции:

- Импорт резюме из внешних источников;
- Отделение резюме от прочих документов;
- Извлечение фактов из текста резюме;
- Классификация резюме;
- Нормализация основных полей;
- Выявление резюме, принадлежащих одному человеку.

Далее кратко рассматриваются наиболее интересные элементы системы.

2 Извлечение фактов

Главным элементом рассматриваемой системы является модуль извлечения фактов. Система обеспечивает выделение следующих полей:

- Личная информация (ФИО, пол, дата рождения/возраст, семейное положение, гражданство);
- Контактная информация (телефоны, электронная почта, ICQ);
- Желаемая должность и пожелания по уровню дохода;
- Опыт работы (период работы, название компании, сфера деятельности компании, должность, отдел, обязанности, стаж);
- Образование (год выпуска, название учебного учреждения, факультет, кафедра, специальность, специализация, квалификация);
- Владение иностранными языками;
- Знание программного обеспечения и технологий;
- Личные качества, хобби;
- Рекомендации (ФИО, должность, название компании, контакты).

Процесс извлечения фактов в описываемой системе состоит из следующих этапов:

1. Графематический анализ. Выполняется разбиение текста на слова и предложения. При этом учитываются характерные для процесса пересылки текста по электронной почте случаи потери форматирования и как следствие – склеивание слов и строк.
2. Морфологический анализ. Морфологический анализ выполняется с помощью набора правил словообразования и словаря лемм-исключений, которые не могут быть описаны набором правил. Важной особенностью данного этапа состоит в том, что язык резюме в значительной мере отличается от языка классических корпусов русского языка (имеет место широкое использование профессиональной терминологии).
3. Выделение лингвистических конструкций вокруг ключевых слов. Иначе говоря, выделяются цепочки слов, обозначающие некоторые объекты предметной области. Например, на основе ключевого слова «университет» будет, в частности, выделено словосочетание «московский государственный университет».
4. В зависимости от контекста, выделенная цепочка может быть интерпретирована как факт. Например, упоминание московского государственного университета в контексте подраздела «образование» означает, что человек учился в данном образовательном учреждении. В контексте подраздела «опыт работы» данное словосочетание интерпретируется как название места работы.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2008, Дубна, Россия, 2008.

3 Классификация резюме

Под классификацией (рубрикацией) резюме понимается его отнесение к одной или нескольким категориям, описывающим профессиональную область или специализацию соискателя. Другими словами, для резюме определяется некая общая характеристика, описывающая соискателя в целом (например, «продажи», «IT-специалисты», «маркетинг»).

Единого стандарта классификации резюме не существует. На практике разные кадровые агентства и job-сайты используют свои собственные схемы классификации. Встречаются как простые одноуровневые классификаторы, так и разветвленные таксономии. Как правило, количество классов верхнего уровня находится в пределах от 10 до 30. Иногда могут применяться схемы, основанные на тегах.

В настоящее время классификатор сайта hh.ru имеет 2 уровня. Каждое резюме может относиться только к одной профессиональной области (верхний уровень) и к нескольким специализациям.

В системе Resumagic была реализована функция автоматической классификации резюме в соответствии со стандартом HeadHunter. Для обучения и настройки автоматического классификатора использовалась выборка из 5661 резюме. Для оценки качества классификации использовалась тестовая выборка из 1887 резюме. Нами были произведены эксперименты классификацией резюме с помощью метода k ближайших соседей и метода PrTFIDF.

3.1 Классификация резюме с помощью метода k ближайших соседей

Метод k ближайших соседей (kNN, k nearest neighbors) заключается в поиске в обучающей выборке документов, наиболее похожих на анализируемый документ, после чего этому документу присваивается самая распространенная категория среди найденных.

Для поиска ближайших соседей вводится метрика близости документов. Каждый документ рассматривается как вектор в n-мерном пространстве, где n – общее количество признаков (термов). В качестве элементов вектора используется мера TF-IDF. В таком представлении метрика схожести 2-х документов определяется как значение косинуса угла между их векторами:

$$Similarity = \frac{\sum_{i=1}^{n} a_i * b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} * \sqrt{\sum_{i=1}^{n} b_i^2}}$$

где a_i и b_i – элементы векторов сравниваемых документов, т.е. TF-IDF.

$$Tfidf = TermFreq * \log\left(\frac{Ndoc}{DocFreq}\right)$$

где

$TermFreq$ – частота встречаемости терма в документе

$Ndoc$ – количество документов в коллекции

$DocFreq$ – количество документов, в которых встречается терм.

Для оценки эффективности нашей реализации метода kNN был произведен ряд экспериментов на тестовом наборе данных. В качестве основного критерия оценки использовалась метрика F1.

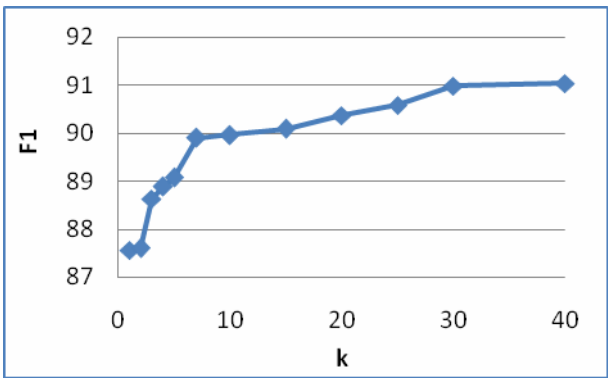

График 1. Зависимость качества классификации резюме (F1) от количества ближайших соседей, принимающих участие в классификации (k).

Таблица 1. Точность, полнота и F1 в зависимости от k

k	Precision	Recall	F1
1	0,778	1	0,876
2	0,778	1	0,876
3	0,796	1	0,886
4	0,800	1	0,889
5	0,803	1	0,890
7	0,816	1	0,899
10	0,817	1	0,899
15	0,819	1	0,901
20	0,824	1	0,904
25	0,828	1	0,906
30	0,834	1	0,909
40	0,835	1	0,910

Как видим из таблицы 1, с помощью нашей реализации метода kNN достигается точность классификации резюме 0,835 при полноте 1, что можно считать удовлетворительным уровнем при автоматической обработке резюме.

3.2 Классификация резюме с помощью метода PrTFIDF

Метод PrTFIDF описан в работе [2]. Формула вычисления наиболее вероятной рубрики выглядит следующим образом:

$$H(d') = \underset{C_j \in C}{\operatorname{argmax}} \sum_{\omega \in \Omega} \frac{Pr(\omega|C_j) * Pr(C_j)}{\sum_{C' \in C} Pr(\omega|C') * Pr(C')} * Pr(\omega|d')$$

где

$Pr(\omega|C_j)$ – средняя частота терма ω в документах категории C_j;

$Pr(C_j)$ – соотношение количества документов в категории C_j к количеству документов в обучающей выборке;

$Pr(\omega|d')$ – частота терма ω в классифицируемом документе.

Метод PrTFIDF также был реализован в рамках платформы Resumagic. Было произведено сравнение результатов работы метода PrTFIDF с результатами kNN. Анализ прогонов на тестовой выборке резюме показал, что при оптимальном количестве используемых для классификации термов (около 1000) методы показывают близкие значения полноты и точности, при небольшом преимуществе метода PrTFIDF.

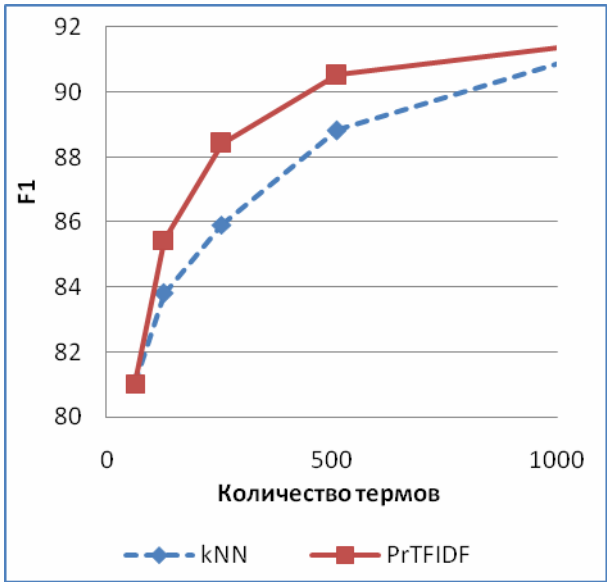

График 2. Сравнение kNN и PrTFIDF

В связи с этим в качестве основного метода автоматического определения профессиональной области в платформе Resumagic был выбран PrTFIDF.

3.3 Выбор термов

Важным фактором, влияющим на качество классификации резюме, является алгоритм выбора термов. Для уменьшения размерности пространства признаков обычно используются два подхода:

- отбрасывание неважных термов;
- кластеризация термов.

При отбрасывании термов для каждого из них рассчитывается некий вес, соответствующий их «важности». Отбрасываются все термы, вес которых ниже определенного уровня. Веса термов могут рассчитываться по-разному. Например, в качестве веса можно брать частоту терма. В Resumagic для вычисления веса используется следующая эвристическая формула:

$$W(\omega) = Freq(\omega|D) * \sum_{C_j \in C} \left(\frac{Freq(\omega|C_j) * N}{Freq(\omega|D)} - 1 \right)^2$$

где

$Freq(\omega|D)$ – частота терма ω в обучающей выборке;

$Freq(\omega|C_j)$ – частота терма ω в категории C_j;

N – количество категорий.

Эксперименты на тестовой выборке резюме показали, что предложенная формула позволяет более точно оценивать важность терма, чем простая оценка частоты. Особенно это заметно при небольшом количестве термов.

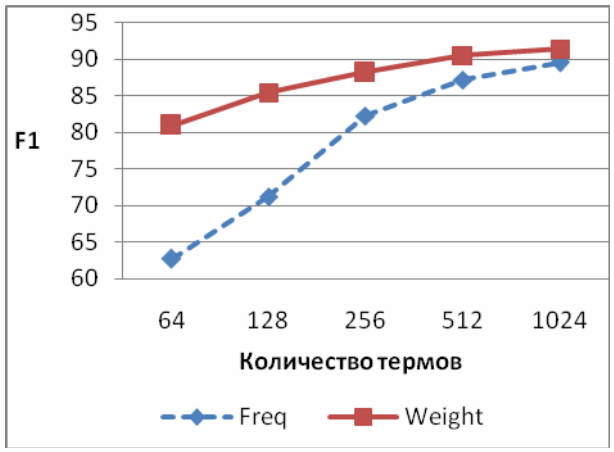

График 3. Сравнение методов выбора термов

Также для уменьшения размерности пространства признаков в Resumagic используется лемматизация. Как показали наши эксперименты, использование морфологии позволяет несколько увеличить качество классификации резюме.

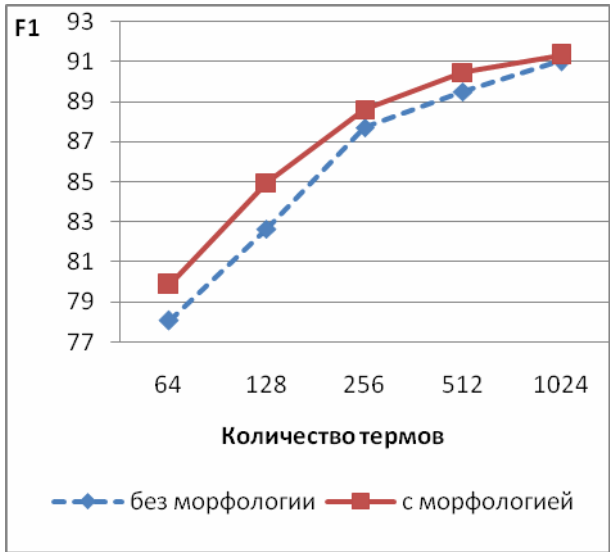

График 4. Сравнение качества классификации с учетом морфологии и без нее.

Использование в качестве термов не только отдельных слов, но и словосочетаний приводит к дальнейшему улучшению качества классификации резюме. Очевидно, что иногда роль слова с точки зрения классификации может сильно изменяться в зависимости от того, в каком контексте это слово употребляется. Например, словосочетания «продвинутый пользователь» и «поддержка пользователей» характерны для совсем разных категорий резюме. Поэтому в платформе Resumagic в качестве термов используются не только отдельные слова, но и словосочетания, что позволяет поднять качество классификации.

3.4 Классификация на основе составленных вручную правил

Помимо метода PrTFIDF для классификации резюме в Resumagic применяется классификатор на основе инженерного подхода. Его работа основана на множестве составленных вручную правил. Инженерный классификатор работает уже не с исходным текстом резюме, а с его структурированным представлением. Можно сказать, что этот классификатор в известном смысле воспроизводит логику рекрутера, занимающегося ручной классификацией резюме. Инженерный классификатор обладает меньшей полнотой (recall), чем PrTFIDF, но при этом обеспечивает почти 100% точность (precision).

Сочетание классификатора на основе правил и метода PrTFIDF обеспечивает распознавания профессиональной области с точностью порядка 0,92 при полноте 0,99.

4 Отделение резюме от прочих документов

Чтобы отличать резюме от других документов, Resumagic рассчитывает специальный числовой индекс под названием ResumagicRank. Этот индекс представляет собой число в диапазоне от 0 до 100. Чем выше ResumagicRank, тем больше текст похож на резюме.

ResumagicRank представляет собой интегральную эвристическую характеристику, которая формируется на основе множества различных факторов. Логику ее расчета можно схематически описать следующей формулой:

$$ResumagicRank = \log\left(\sum_{i}^{NP} WP_i * \sum_{i}^{NB} WB_i\right)$$

где

NP – количество фактов личного характера, извлеченных из документа. Под «личными фактами» подразумевается непосредственная информация о человеке (ФИО, возраст, семейное положение и т.д.).

WP_i – вес (значимость) i-того личного факта.

NB – количество фактов делового характера, извлеченных из документа. Под деловыми фактами подразумевается информация, связанная с профессиональной деятельностью человека, его умениями и знаниями (опыт работы, образование, желаемая должность, желаемая зарплата и т.д.).

WB_i – вес (значимость) i-того «делового» факта.

Как видно из формулы, важным требованием к документу является одновременное присутствие как деловой, так и личной информации.

Наши эксперименты показали, что ResumagicRank позволяет с высокой точностью отличать резюме не только от обычного спама, но и от «резюмеподобных» писем. Например, в обычной рабочей переписке может содержаться достаточно много информации, характерной для резюме (хотя бы подписи к письму). Также некоторые информационные рассылки могут включать контактную информацию, сведения о людях, компаниях и т.п.

ResumagicRank используется для сортировки входящих документов по 3 разделам: «резюме», «не резюме» и «нераспознанные». Документы, которым присвоен ResumagicRank > 66, относятся к категории «резюме». Если ResumagicRank < 33, то документ помечается как «не резюме». Остальные документы относятся к классу нераспознанных и нуждаются в дальнейшей ручной проверке. Как правило, в «нераспознанные» попадают резюме, составленные в нетрадиционной манере или содержащие слишком мало сведений. Также к категории «нераспознанных» могут быть отнесены некоторые письма, содержащие большое количество фактов о людях (например, новостные рассылки с информацией о назначениях и отставках в правительстве).

5 Определение региона

Город проживания не всегда указывается в резюме в явном виде. Кроме того, иногда рекрутера могут интересовать кандидаты не из конкретного города, а из достаточно большого списка разных городов (например, ближнего Подмосковья).

Для распознавания региона в Resumagic имеется специальный механизм, работающий по следующему алгоритму:

1. Поиск в тексте резюме непосредственного упоминания региона. Иногда область, в которой проживает соискатель, указывается явным образом. Например, в некоторых резюме соискатели пишут что-то вроде этого: «регион проживания – Краснодарский край».
2. Если регион не указан явно, Resumagic пытается определить его на основании города, в котором проживает соискатель. Для этого имеется внутренний справочник соответствий между городами и регионами. С помощью этого справочника, например, можно определить, что город Борисоглебск относится к Воронежской области. На этом этапе проблема может состоять в том, что и город не всегда указывается кандидатом в своем резюме. Если город не указывается в явном виде, то для его определения используются дополнительные проверки:
   a. Код города в телефоне. В Resumagic имеется справочник телефонных кодов для основных городов России.
   b. Станция метро. Иногда в резюме отсутствует прямая информация о городе, но упоминается ближайшая станция метро. Особенно это характерно для тех резюме, которые соискатели отправляют конкретному работодателю или в конкретное кадровое агентство, предполагая, что город и так очевиден. Многие названия станций метро являются «эндемичными», т.е. встречаются только в одном городе.

3. В разных регионах встречаются города с одинаковыми названиями. Однако при указании почтового адреса в резюме иногда указывается индекс, что позволяет надежно определить регион.

6 Защита от дублирования

При поступлении новых резюме необходимо следить за тем, чтобы в базе данных не возникало дублирования соискателей. Ручная проверка на дублирование отнимает некоторое время. При этом, как правило, в базе все равно постепенно появляются «двойники». Если одному соискателю соответствует несколько разных записей, то это может привести к путанице и снижению эффективности работы рекрутера.

Для решения этой проблемы в описываемой системе реализована автоматическая защита от дублирования.

Защита от дублирования включает в себя следующие уровни:

1. Поиск по полному совпадению текстов резюме на основе хеш сумм. После очистки от элементов форматировании для входящего резюме вычисляется CRC32, а затем в базе находятся все резюме с совпадающим значением CRC32. Найденные резюме проходят дополнительное проверочное сравнение с входящим документом, в результате чего происходит обнаружение полных дубликатов. Эта проверка позволяет защититься в случаях, когда соискатель присылает одно и то же резюме по нескольку раз.
2. Поиск по совпадению ФИО, даты рождения или контактной информации. Если из текста входящего резюме были успешно выделены фамилия, имя, отчество, дата рождения или контактная информация, то Resumagic производит поиск соискателя по этим полям. При этом принимаются во внимание также распространенность фамилии и имени соискателя.
3. Поиск частичных совпадений резюме. Если предыдущие шаги не дали результатов, то система пытается принять решение с помощью «индекса сходства резюме». Индекс сходства – это число от 0 до 100, которое показывает, насколько два резюме похожи между собой. Эта методика позволяет найти дублирующихся соискателей, даже если они внесли изменения в текст своего резюме. Для этого сначала в базе данных производится поиск соискателей только по имени и фамилии или по контактной информации. Потом для всех найденных резюме определяется степень их похожести на входящее резюме. Если система находит в базе текст с высоким индексом сходства, то это означает, что кандидат прислал обновленное резюме с некоторыми изменениями. В этом случае можно уверенно обновлять соискателя с максимальным индексом схожести. Если же индекс схожести низок, то это говорит о том, что резюме принадлежат разным лицам. Тогда можно спокойно создавать нового соискателя, не опасаясь дублирования.

Литература

[1] Resumagic. http://resumagic.ru

[2] Joahims. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF. Proceedings of ICML-97, 14th International Conference on Machine Learning. http://www.cs.cornell.edu/people/tj/publications/joachims_97a.pdf

[3] Гершензон. Технология извлечения структурированной информации из неструктурированного текстового массива. 2006. Jug.ru. http://www.jug.ru/servlets/images/meeting_2006_12_23/FactExTech.ppt

[4] Качаева, Южиков. Автоматизированная система распознавания и классификации резюме.RCDL-2007. http://rcdl2007.pereslavl.ru/papers/paper_45_v1.pdf

[5] Петров, Кузнецов. Особенности сетевого англоязычного лингвистического процессора для формализации текстовой информации на естественном языке. Диалог-2006. http://www.dialog-21.ru/dialog2006/materials/html/Petrov.htm

[6] Resume Mirror. http://www.resumemirror.com/products/resume-processing.html

Resumagic: automatic CV processing

A.V. Safronov

This paper describes Resumagic, a system intended for automated parsing of curriculum vitae texts [CVs]. Problems of filtering and classification of resumes were reviewed. Also results of applied methods are given.