УДК 004.021

Обзор алгоритмов автоматического извлечения ключевых слов из текста

И.А. Резников
Московский государственный университет им. М.В. Ломоносова
Вычислительный центр им. А.А. Дородницына РАН

Ключевые слова зачастую используются поисковыми движками и библиотеками документов для поиска информации и определения сходства нескольких разных текстов. Прочтение и аннотирование содержания больших текстов – сложная для человека задача, настолько, что её выполнение ограниченным числом людей в условиях постоянного поступления новой информации практическим невозможно. В результате для решения этой задачи всё чаще применятся автоматические системы. Сложность задачи состоит, во-первых, в сложности обработки естественного языка, а также в определении, насколько хорошо слово или группа слов представляют темы из исходного текста. Развитие интернета принесло одновременно и большое количество информации для анализа, и потребность в поиске по этой информации. Извлечение ключевых слов – метод, используемый поисковыми движками и индексаторами для быстрого определения темы текста и нахождения в нем конкретной информации.

Для решения этой задачи существует уже множество методов, и постоянно разрабатываются новые решения. Несмотря на разницу в методах, большая их часть пытается сделать примерно одно и то же: используя некоторую эвристику (например, расстояние между словами, частоту использования слов или заранее определённые отношения между словами), найти группу слов, которая точно определяет темы или описывает информацию, содержащуюся в исходном тексте.

Анализ частоты слов

Многие ранние работы концентрировались на частоте использования термина в тексте, но большая их часть рассматривала ключевые слова в отношении только одного документа. Только в 70-х годах стала более популярна идея статистически анализировать частоту употребления слов в документе по отношению к большому количеству других документов. Этот приём, известный как Term Frequency - Inverse Document Frequency, или просто как TF-IDF, даёт вес каждому термину в соответствии с тем, насколько хорошо данный термин определяет документ по отношению к корпусу (собранию документов). Для этого термину добавляется вес за количество раз, которое он появляется в данном документе, и убавляется за количество других документов, в которых этот термин также

употребляется. Рассмотрим термин t и документ d ∈ D, где t появляется в n из N документов корпуса D. TF_IDF принимает следующую форму:

$$TFIDF(t,d,n,N) = TF(t,d) \times IDF(n,N) \quad (1)$$

Существует множество возможных функций TF и IDF. На практике для TD и IDF может быть использована практически любая функция. Часто используются следующие:

$$TF(t,d) = \begin{cases} 1 \text{ if } t \in d \\ 0 \text{ else} \end{cases} \quad (2)$$

$$TF(t,d) = \sum_{word \in d} \begin{matrix} 1 \text{ if word} = t \\ 0 \text{ else} \end{matrix} \quad (3)$$

В добавок к этому, частота термина может быть нормализована в некотором диапазоне. Затем она комбинируется с IDF функцией. Примеры IDF функций:

$$IDF(n,N) = \log\left(\frac{N}{n}\right) \quad (4)$$

$$IDF(n,N) = \log\left(\frac{N-n}{n}\right) \quad (5)$$

Таким образом, примером возможной функции TF-IDF может быть:

$$TFIDF(t,d,n,N) = \left(\sum_{word \in d} \begin{matrix} 1 \text{ if } word = t \\ 0 \text{ else} \end{matrix}\right) \times \log\left(\frac{N-n}{n}\right) \quad (6)$$

Когда TF-IDF функция применена ко всем словам во всех документах корпуса, слова можно отсортировать во полученным весам. Более высокий TF-IDF вес говорит о том, что слово важно для данного документа, и в то же время достаточно редко употребляется в других документах корпуса. Это зачастую можно интерпретировать как знак того, что слово является важным для данного конкретного документа и может быть использовано, чтобы точно описать документ. TF-IDF предоставляет хорошую эвристику для определения кандидатов в ключевые слова, и этот метод (и многие его модификации) за годы исследований показал свою эффективность. Со времени публикации метода в 1972 году, было разработано множество новых методов извлечения ключевых слов, и многие из

них по-прежнему полагаются на TF-IDF. В связи с эффективностью и простотой, TF-IDF продолжает активно применяться и сегодня.

Частота совместного появления слов

В то время как многие методы извлечения ключевых слов полагаются на частоту встречаемости слова (в документе, в корпусе, или их комбинации), у метрик такого рода существуют и некоторые проблемы [7] [4], такие как зависимость от корпуса и опора на предположение, что точное ключевое слово должно часто встречаться в документе, но редко в других документах корпуса. Эти методы так же игнорируют любые возможные отношения между словами в документе.

Использование корпуса документов

Одна из попыток использовать эту информацию использует Цепь Маркова, которая применяется для оценки каждого слова в корпусе из всех документов. Этот метод определяет Цепь Маркова для документа d и термина t с двумя состояниями (C, T). Вероятность перехода из C в T – это вероятность, что из всех документов данный термин встретился именно в документе d (в сущности число раз, которое t встретился в d, делённое на число раз, которое t встретился во всех документах). Вероятность перехода из T в C – это вероятность, что из всех терминов в d встретился именно этот (число раз, которое t встретился в d, делённое на число встреченных в d терминов). Идея метода заключается в том, что если два термина переходят в одно состояние с близкой вероятностью, то они связаны.

Авторы метода определили, что слово имеет меньшую вероятность описывать документ, если оно приходит в некоторое состояние с частотой, близкой к аналогичной частоте других слов в документе (названой фоновым определением), и большую вероятность описывать документ, если оно сильно отличается от этого распределения. Этот метод сопоставим, и зачастую превосходит TF-IDF по точности [7].

Извлечение ключевых слов из одного документа на основе частоты слов

Большая часть методов извлечения ключевых слов полагается на сравнение документа с корпусом для определения наиболее уникальных слов для конкретного документа. Однако такую меру сложно использовать с маленьким корпусом, и тем более при его отсутствии.

Один из методов извлечения ключевых слов без корпуса документов заключается в построении матрицы совместного появления слов, как на Таблице 1 [5]. Совместным

появлением двух слов считается их присутствие в одном предложении. В приведённом примере, слова b и c встречаются в одном предложении 42 раза.

	a	*b*	*c*	*d*
a		5	13	7
b	5		42	3
c	13	42		25
d	7	3	25	

Таблица 1: Пример матрицы совместного появления

Авторы утверждают, что слова более важны для документа, если они чаще встречаются вместе с другими словами документа. Для некоторого слова w_i это можно представить, как отношение совместных появлений слов w_i и w_j к числу всех совместных появлений пар слов с участием w_i. С таким предположением, большее значение данного отношения будет означать, что слово w_i скорее всего является ключевым словом документа.

Проблемой является случай, когда слово встречается в документе всего несколько раз, так как отношение в этом случае основано на статистически неточной информации. Для предотвращения этого, авторы используют критерий согласия Пирсона для каждого слова в документе.

$$X^2 = \sum_{i=0}^{n} \left(\frac{(O_i - F_i)^2}{E_i} \right) \quad (7)$$

где n – число слов в документе, O – наблюдаемая частота и E – ожидаемая частота. Этот критерий позволяет замерить и сравнить частоту распределения каждого слова с ожидаемой. Слово, которое встречается малое число раз, будет иметь распределение, близкое к ожидаемому случайному, и низкое значение X^2, а часто встречающееся слово, которое также часто встречается вместе с другими словами, будет иметь высокое значение X^2.

Авторы показали, что данный метод схож по точности с TF-IDF, но не требует корпуса документов.

Извлечение ключевых слов, чувствительное к содержимому

Ещё один метод извлечения ключевых слов [5] подходит к проблеме с другой стороны. В то время, как многие способы извлечения ключевых слов полагаются на статистическую информацию по частоте встречаемости слов в документе, этот метод, названый KeyGraph,

вместо этого полагается на кластеризацию взаимосвязанных терминов в группы, чтобы определить важные для содержания документа слова.

KeyGraph строит представление документа в виде графа, в котором вершинами являются термины, а ребрами – часто встречающиеся совместные появления терминов в тексте. Затем с помощью нахождения максимально соединённых подграфов выделяются кластеры терминов. После этого кандидаты в ключевые слова выбираются как вершины, которые имеют инцидентные рёбра, соединяющие два различных кластера. Интуитивно, эти кандидаты являются терминами, соединяющими различные идеи или концепты (кластеры) документа. Выбранные кандидаты ранжируются по вероятности того, что для двух кластеров, в которых они состоят, они являются словами, соединяющими кластеры (слова, чаще всего использованные для объединения этих кластеров).

Тесты KeyGraph показывают, что он сопоставим и даже превосходит TF-IDF [5]. В добавок была проведена серия тестов на данных из социальных сетей во время президентских выборов в 2008 году, которая показала, что KeyGraph способен найти ключевые слова в зашумлённой информации с большим количеством не имеющих отношения к теме данных.

Извлечение ключевых слов с использованием лексических цепей

Лексические цепи – это просто список связанных слов из текста. Отношения между ними обычно семантические, такие как синонимия или гиперонимия. Примером лексической цепи может служить следующее:

сахарный клён → клён → дерево → растение (8)

Такое представление семантической информации из естественного языка позволяет закодировать в цепь контекст употребления слов. В приведённом примере можно увидеть, что слово «растение» следует за слово «дерево». Слово «растение» следует за «дерево» так как «дерево» связано со словом «клён». Если бы не эта семантическая информация, следующее слово в цепи могло бы быть чем-то связанным с графами или структурами данных.

Лексические цепи довольно часто находят своё применение в методах автоматического реферирования текстов [1], где с помощью них можно быстро и точно находить термины со схожими значениями. Рассмотрим пример: «Сахарный клён – это клён, который…» По лексической цепи примера можно понять, что «сахарный клён» и «клён» имеют очень

схожее значение, и одни из них может быть кандидатом на удаление из текста для более краткой аннотации фразы.

Такие лексические цепи могут быть использованы для нахождения важных для текста слов [2]. Для этого используется статистический классификатор, который строит деревья принятия решений чтобы определить, что данное слово имеет высокую вероятность быть ключевым. Для этого каждому термин из текста относят к некоторой лексической цепи, которая его содержит. Затем терминам присваивается вес, зависящий от того, когда он встретился первый и последний раз, от частоты встречаемости термина и от первых и последних положений в документе синонимов и гиперонимов термина. При использовании этого метода авторами была достигнута точность 64%.

Извлечение ключевых слов с помощью Байесовского классификатора

Ещё одним предложенным для извлечения ключевых слов из документа в составе корпуса методом был адаптированный TF-IDF совмещённый с наивным Байесовским классификатором. Этот метод применяет формулу 9 для каждой фразы в документе.

$$TFIDF(p,d) = \mathrm{P}[phrase \text{ in } d \text{ is } p] \times (-\log \Pr[p \text{ appears in any document}]) \quad (9)$$

где p – рассматриваемая фраза и d – текущий документ. Вероятность, что фраза является ключевой затем определяется по Теореме Байеса:

$$\Pr[key \mid T, D] = \frac{\Pr[T \mid key] \times \Pr[D \mid key] \times \Pr[key]}{\Pr[T, D]} \quad (10)$$

Где T это TD-IDF значение, вычисленное ранее, и D – расстояние от начала документа до первого появления данной фразы (количество фраз перед ней). Таким образом, Pr[T | key] – это вероятность того, что рассматриваемая фраза имеет TF-IDF значение T, Pr[D | key] – вероятность того, что фраза появляется в документе на расстоянии D и Pr[key] – это вероятность, что из всех фраз в документе эта является ключевой. Pr[T, D] используется для нормализации результатов в диапазон [0, 1].

После этого фразы сортируются по вероятности быть ключевыми с учётом T и D, и k ключевых фраз извлекаются из верхушки этого отсортированного списка.

Авторы показали, что этот метод даёт схожие или слегка лучшие результаты, чем TF-IDF [3].

Заключение

TF-IDF – один из наиболее известных алгоритмов для извлечения ключевых слов, которые используются в настоящее время [6] при доступном корпусе документов. Некоторые более новые методы используют TF-IDF как часть вычислительного процесса и многие другие полагаются на ту же основу, что и TF-IDF. Практически все методы, использующие корпус, зависят от взвешивающей функции, которая балансирует некоторые свойства термина или фразы в конкретном тексте (частоту, позицию в документе, совместное появление с другими словами) со схожими мерами для всего корпуса.

При отсутствии корпуса алгоритмы извлечения ключевых слов нуждаются в дополнительных свойствах слов, по сравнению с используемыми TF-IDF и схожими методами. Эта дополнительная информация может быть получена лексическим или семантическим анализом, или какой-нибудь мерой совместной встречаемости слов.

Литература

1. *Barzilay R., [at al]*. Using lexical chains for text summarization / In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization – 1997. - V. 17. - P. 10–17.
2. *Ercan G., Cicekli I.* Using lexical chains for keyword extraction // Information Processing & Management – 2007. - № 43. – P. 1705–1714.
3. *Eibe F., [at al]*. Domain-specific keyphrase extraction // In Proceedings of Seventeenth International Joint Conference on Artificial Intelligence - 1999. – P. 668.
4. *Matsuo Y., Ishizuka M.* Keyword extraction from a single document using word co-occurrence statistical information - International Journal on Artificial Intelligence Tool – 2004. – P. 13.
5. *Ohsawa Y., Nels E., Masahiko Y.* KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor // In Proceedings of the Advances in Digital Libraries Conference - 1998 - P. 12.
6. *Robertson S.* Understanding inverse document frequency: on theoretical arguments for idf - Journal of Documentation – 2004. – № 60. – P. 503–520.
7. *Wartena C., Brussee R., Slakhorst W.* Keyword extraction using word co-occurrence // In Proceedings of the 2010 Workshops on Database and Expert Systems Applications – 2010. – P. 54–58.