G. G. Novikov, I. M. Yadykin

Geometrical Fuzzy Search Method for the Business Information Security Systems

Keywords: information security system, fuzzy search, geometrical fuzzy search method, full text search

The main reason of the article is how to use one of new fuzzy search method for information security of business or some other purposes. So many sensitive information leaks are through non-classified documents legal publishing. That's why many intelligence services like to use the "mosaic" information collection method so much: This article is about how to prevent it.

Г. Г. Новиков, И. М. Ядыкин

ГЕОМЕТРИЧЕСКИЙ ПОДХОД К ПРОБЛЕМЕ НЕЧЕТКОГО ПОИСКА ФРАГМЕНТА ТЕКСТА В КОНТУРЕ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ ПРЕДПРИЯТИЯ

Введение

Настоящая работа посвящена разработке подхода к задаче анализа текстов, использующихся при открытом документообороте предприятия. Как известно из практики разведывательной деятельности и промышленного шпионажа во всех странах, большой процент полезной информации получается в результате просеивания публикаций в открытой печати и составления информационной мозаики, позволяющей судить о конфиденциальном содержимом без его получения в явном виде.

В этом контексте весьма продуктивным видится поход к обеспечению безопасности конфиденциальных данных предприятия, при котором контрмеры принимаются путем включения в контур информационной безопасности предприятия системы нечеткого полнотекстового поиска. В задачу такой системы входит выходной контроль открытой текстовой документации, легально выходящей за пределы контура информационной безопасности предприятия, предположительно не содержащей конфиденциальных сведений.

Метод выходного контроля информации

Исходными данными для такой системы являются тексты документов, отнесенные к конфиденциальным сведениям и находящиеся в информационной базе на защищенных серверах внутри контура информационной безопасности предприятия. Любой исходящий документ, покидающий контур информационной безопасности, должен пройти анализ, при котором определяется степень его подобия с любым из документов, находящихся в базе конфиденциальных. Результат анализа, выраженный в количественной мере, называемой релевантностью, указывает специалисту по информационной безопасности предприятия на возможную потенциальную утечку конфиденциальных сведений.

Ключевым моментом в этом процессе является сравнение двух текстов.

Понятие «сравнение» в контексте поставленной задачи сводится к отысканию похожих документов или фрагментов документов с количественной оценкой степени этой похожести — релевантности.

С точки зрения информационных технологий эта задача представляет собой нахождение похожих фрагментов текстов и относится к области нечеткого поиска.

Исследование типов современных поисковых систем и анализ современного состояния разработок [1—3] показали, что на фоне достаточно широкого разнообразия методов нечеткого поиска существует ряд проблем, стоящих на пути решения этой задачи:

- во-первых, на данный момент в «мире поиска» наблюдается явное доминирование именно фразовых поисковиков как в корпоративном секторе, так и в глобальном интернет-хаосе; иначе говоря, запросом является одно или несколько слов — короткая фраза;
- во-вторых, в явном виде ни один из опубликованных методов нечеткого поиска не подходит для решения поставленной задачи, а именно нахождения похожих фрагментов текстовых документов с количественной оценкой степени похожести во временных рамках оперативного документооборота.

Предлагаемая методика количественного определения степени похожести документов, будем называть ее релевантностью, заключается в следующем.

Все множество текстовых документов поисковой базы рассматривается как ряд своеобразных графических изображений, в которых пикселем, или минимальным графическим примитивом, является одно слово документа. Таким образом, результатом сравнения является совокупность совпадений слов в исследуемом тексте документа и поисковом шаблоне с учетом их взаимного (геометрического) расположения на странице текста, указывающего на плотность таких совпадений.

Места концентрации совпавших слов будут частными вариантами результата поиска по сходству. Они будут состоять из цепочки слов, то есть последовательности слов запроса, расположенных друг от друга на расстоянии не более некоторой задаваемой константы, регулирующей допустимую к рассмотрению степень подобия. Релевантность, рассчитанная для каждой из таких цепочек в пределах анализируемой части сравниваемых документов, дает аддитивный вклад в общую оценку релевантности сравниваемых текстов.

Релевантность, или степень близости информационного содержания сравниваемого документа из базы и документа запроса (подобие поискового запроса), будет определяться двумя параметрами:
- количественным, то есть количеством слов поискового шаблона, содержащихся в цепочке Qnt (от англ. quantity — количество),
- качественным, то есть расстоянием найденных слов по отношению друг к другу или их общей плотностью Qlt (от англ. quality — качество).

Расчет количественного значения релевантности в системе производится по формуле:

$$R = Qnt * Qlt \quad .$$

Или более развернуто:

$$R = \frac{M}{N} \cdot \frac{\sum_{i=1}^{M-1} \left[\left(Lim - E_{\min_i} \right) + 1 \right]}{Lim(M-1)} \quad ,$$

где M — количество уникальных слов, содержащихся в найденном текстовом фрагменте; N — общее количество уникальных слов запроса; Lim — максимальная удаленность слов друг от друга, при которой происходит их включение в результат; E_{mini} — расстояние двух соседних слов результата.

При разработке формулы релевантности учитывались следующие соображения.

В случае полного совпадения запроса и найденного текстового фрагмента документа поисковой базы релевантность принимает значение 1, а в случае отрицательного результата поиска — 0. Исходя из этого, значения параметров *Qlt* и *Qnt* также варьируются от 0 до 1. *Qnt* равен 1, если все слова запроса присутствуют в результате, а *Qlt* равен 1, если все найденные слова запроса расположены друг от друга на расстоянии не более одного символа.

Поиск производится по индексированному файлу, назовем его словарем, полученному в результате обработки информационной базы системы. В процессе создания результативных цепочек слов участвуют только слова запроса, длина которых больше или равна трем символам и которые содержатся в специальном файле, назовем его подсловарем, полученном путем обработки запроса.

Алгоритм построения результативных цепочек слов. Алгоритм достаточно сложный и предполагает множество проверок и учет различных особенностей, влияющих на его оптимальность и корректность вычислений. Часть разработанного алгоритма реализована в программе нечеткого поиска, зарегистрированной в государственном реестре № 2014612476 от 26.02.2013. На остальные части поданы заявки на регистрацию.

Содержимое подсловаря, участвующее в формировании результата. Это некоторая структура, состоящая из множества позиций вхождения слов запроса в поисковую базу (понятно, что одно слово может встречаться несколько раз). Здесь в каждой строке расположены позиции вхождения одного слова. Строки следуют в порядке, определенном последовательностью слов запроса. При значении $Lim = 6$, а $M = 5$ строятся три цепочки, для которых значения релевантности различаются — чем больше слов запроса найдено и чем геометрически ближе они расположены друг к другу, тем релевантность выше.

На базе разработанного поискового метода была создана программная система, обладающая рядом интересных функциональных особенностей, а именно:

- инвариантностью по отношению к тексту (не зависит от структуры и форматирования);
- лингвонезависимостью (применим к текстам на разных языках);
- отсутствием зависимости от длины запроса и объема исходных данных;
- отсутствием зависимости от смысловой точности запроса (не нужно конкретизировать запрос для получения требуемых результатов).

Заключение

Полученные результаты могут быть использованы в любых других областях применения, связанных с нечетким поиском текстовой информации, например:

- выявление фактов плагиата (отслеживание заимствования чужой информации без уведомления автора, например, в рефератах, докладах, научных работах и пр.);
- организация полнотекстового поиска в пределах корпоративной сети с целью консолидации данных;
- организация поиска на веб-сайтах по заранее созданной базе ресурсов (использование в качестве внутреннего поискового модуля).

СПИСОК ЛИТЕРАТУРЫ:

1. *Sebastiani F.* Machine learning in automated text categorization // ACM Computing Surveys. 2002. Vol. 34. № 1. P. 1–47.
2. *Chomsky N.* Syntactic Structures. The Hague: Mouton, 1957.
3. *Manning Ch. D., Schьtze H.* Foundations of Statistical Natural Language Processing. MIT Press, 1999.

REFERENCES:

1. *Sebastiani F.* Machine learning in automated text categorization // ACM Computing Surveys. 2002. Vol. 34. № 1. P. 1–47.
2. *Chomsky N.* Syntactic Structures. The Hague: Mouton, 1957.
3. *Manning Ch. D., Schьtze H.* Foundations of Statistical Natural Language Processing. MIT Press, 1999.