*Проблемы бионики*
*1999. Вып. 51. С. 146–150*
ХТУРЭ
УДК 025.4.03

*Ю.В. ЛАНДГРАФ*

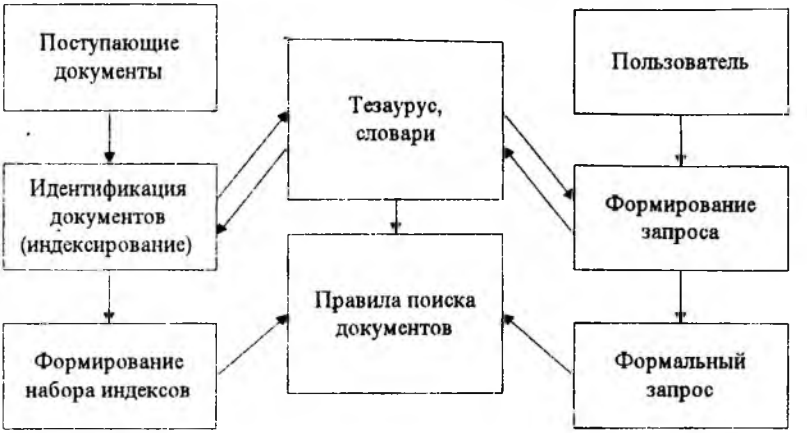
# ТЕХНОЛОГИИ ПОИСКА ЕСТЕСТВЕННО-ЯЗЫКОВОЙ ИНФОРМАЦИИ В СОВРЕМЕННЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ НА ОСНОВЕ СЕМАНТИЧЕСКИХ МОДЕЛЕЙ

Потоки информации, получаемой специалистами в различных областях знаний из разнообразных источников, неуклонно растут. Согласно статистике, общее количество доступной информации удваивается каждые четыре-пять лет. К сожалению, несмотря на почти экспоненциальный рост количества информации, наши возможности по ее поиску и извлечению необходимых сведений, наоборот, уменьшаются.

В связи с бурным развитием компьютерных технологий возросла роль различных информационных систем, особенно информационно-поисковых систем (ИПС), позволяющих осуществлять быстрый поиск информации в глобальных информационных хранилищах данных. Особенно актуально внедрение ИПС в связи с бурным ростом сети Интернет, которая на данном этапе позволяет получить доступ к обширным базам данных и разнообразной информации, расположенной на Web-сайтах практически во всех уголках земного шара. Большая часть этой информации поступает из Интернет, где она существует в электронной форме в виде текстовых документов, что позволяет создавать специальное программное обеспечение для нахождения нужной информации по каким-либо заданным критериям. Среди ИПС наиболее распространены системы, использующие лексические методы поиска, такие, например, как поиск по ключевым словам [1]. Несмотря на очевидные положительные результаты, данный подход не решил проблему поиска информации, так как очень часто пользователь в ответ на запрос получает документы, которые на самом деле иррелевантны к данному запросу, и наоборот, часть релевантных документов попросту не может быть найдена, поскольку они могут не содержать каких-то ключевых слов. В качестве примеров таких систем можно привести очень популярные во всем мире поисковые Интернет системы Infoseek, Yahoo, Lycos и т.д.

Таким образом, на сегодняшний день остается актуальной проблема определения смысла документа, т.е. того, о чем идет речь в документе. Смысловое представление необходимо для того, чтобы определить, какой документ соответствует запросу, а какой нет. Следовательно, современная ИПС должна хранить не только сам документ, но и некую модель, определяющую содержимое документа.

Логическая организация условной информационно-поисковой системы может быть представлена в следующем виде (рисунок):




В соответствии с данной схемой основные характеристики каждого поступающего документа (автор, название, тема) определяются на этапе индексирования. Администратор или система, составляющая индекс, может обращаться к поисковым тезаурусам и/или вспомогательным словарям для того, чтобы назначить соответствующие индексные термины каждому документу. Для каждого документа формируется набор индексов для последующего использования при поиске документов.

Пользователь при формировании запроса также может использовать тезаурус и словари, но поскольку большинство пользователей не имеют представления о том, какие словари они могут использовать для формирования информационно-поисковых запросов, то это приводит к ошибочным результатам, так как словари, используемые пользователем, могут не совпадать со словарями, используемыми системой.

Сам по себе поиск и извлечение информации осуществляется путем сравнения по особым правилам набора индексов и формального запроса. Смысл этих правил – поиска документов – состоит в том, что система должна по запросу выбрать все и только те индексы, которые попадают в подмножество индексов, определяемое запросом.

Таким образом, система извлечения документов включает в себя 3 основные части:

- хранилище документов или их представлений;
- пользователи системы, требующие удовлетворения их информационных потребностей;
- правила поиска документов, которые сравнивают представление документов с представлением пользовательских запросов.

Недостатки в классических ИПС заставили пересмотреть подход к проблеме информационного поиска. Стало очевидным, что большей эффективности можно достигнуть, если перейти от лексического к более перспективному семантическому поиску, тем самым переведя ИПС в класс Интеллектуальных Информационных Систем.

За последние годы были предприняты различные подходы к решению проблемы поиска информации, учитывающие семантику информационных документов. В частности, были разработаны методы, основанные на деревьях решений и индукционных правилах, нелинейной регрессии и классификации, нейронных сетях [2].

В данной статье исследуется еще один метод для поиска информации – метод латентно-семантического индексирования (Latent-Semantic Indexing - LSI). Он является эффективным методом поиска в текстовых документах и использует некоторые принципы вычисления семантического расстояния в текстах [3]. Традиционные лексически ориентированные методы пытаются сопоставлять слова в запросе со словами в документах, что приводит к тому, что извлекаются нерелевантные документы, а часть релевантных документов теряется. Это называется проблемой совпадения слов (word-matching problem), она возникает из-за того, что разные слова могут иметь одинаковые значения (синонимия), в то же время многие слова могут иметь более одного значения (полисемия). Метод LSI преодолевает эту проблему, поскольку использует статистически полученные концептуальные (понятийные) индексы, а не отдельные слова. LSI – это векторно-пространственный подход к смысловому поиску информации. Метод LSI, как и другие векторно-пространственные методы, базируется на предположении, что смысл текста может быть определен на основании терминов, содержащихся в этом тексте. Идея векторно-пространственного подхода состоит в том, что документы (естественно-языковые тексты) представляются векторами терминов

$$d = (t_1, t_2, \ldots, t_n),$$

где $t_i$ $(1 \le i \le n)$ - неотрицательное число, определяющее количество вхождений термина $t_i$ в документ **d**. Таким образом, коллекция документов вместе с содержащимися в них терминами образует пространство термин-документ (term-document space – TDS), определяемое матрицей термин-документ. Каждый элемент этой матрицы представляет собой неотрицательное число, определяющее количество вхождений определенного термина в определенный документ. Размерность этой матрицы зависит от количества документов в коллекции и количества различных терминов в них, и может составлять тысячи строк и столбцов. LSI-метод анализирует матрицу с тем, чтобы выявить скрытые (латентные – отсюда и название метода) ассоциативные связи между терминами и документами. LSI анализирует "похожесть" контекстов, в которых используются те или иные термины, и формирует уменьшенное подпространство, в котором

термины, употребляемые в похожих контекстах, и документы, соответствующие этим контекстам, расположены рядом друг с другом. Для этого LSI использует метод линейной алгебры разложения собственных значений (Singular Value Decomposition – SVD). Данный метод основан на статистических методах (в частности, на факторном анализе) и хорошо описан в литературе [4].

В сформированном подпространстве термины и документы сгруппированы по смыслу, так что его можно назвать семантическим пространством для данного корпуса текстов и входящих в них терминов. Каждому термину и документу в семантическом пространстве соответствует вектор, определяющий положение конкретного термина или документа в данном подпространстве. Если запрос также представить в виде вектора в этом пространстве, то можно отсортировать документы по степени релевантности к запросу, анализируя направления векторов.

Достоинством метода является то, что для выявления ассоциативных связей между терминами не нужно пользоваться никакими словарями, тезаурусами или базами знаний – эти связи определяются на основе автоматической обработки уже существующих текстов.

Метод LSI оценивает семантическое содержание документов и использует эту оценку для определения уровня релевантности документа к запросу пользователя. Поскольку поиск базируется на концепте документа, а не на отдельных терминах, содержащихся в нем, LSI может извлекать документы, релевантные к пользовательскому запросу, даже если запрос и документ не используют никаких общих терминов.

Существует множество дополнений к классическому методу LSI, позволяющих улучшить результаты поиска, увеличить скорость обработки огромных матриц термин-документ, а также применять метод LSI в других областях, связанных с информационными технологиями, например, для автоматического анализа динамически обновляемых баз данных с последующим принятием решений, для перевода текстов с одного языка на другой, и так далее.

Основная задача поиска – найти как можно большее число из всех документов, соответствующих запросу (т.е. релевантных документов), и как можно меньшее количество несоответствующих документов (выбранных ошибочно, т.е. нерелевантных).

Для оценки эффективности систем поиска (извлечения) информации широко используются два основных критерия эффективности поиска - точность (presicion) и выборка (recall).

Точность извлечения документов определяется как отношение количества выбранных релевантных документов к общему количеству выбранных документов. Например, если система по запросу пользователя

находит 3 документа, из которых на самом деле только один релевантен, то точность системы равна 33%.

Оценка выборки более сложна, так как она требует отыскания релевантных документов, которые не были извлечены во время поиска по запросу пользователя. Выборка представляет собой отношение количества извлеченных (найденных) релевантных документов к общему количеству релевантных документов в коллекции. Предположим, что в предыдущем примере в коллекции на самом деле существует 4 релевантных документа, из которых система смогла найти только один. Тем самым выборка будет равна 25%.

Существует еще один параметр, используемый при оценке эффективности систем извлечения информации – выпадение. Выпадение определяется как отношение выбранных нерелевантных документов к общему количеству нерелевантных документов в коллекции. Предположим теперь, что в предыдущем примере всего 14 документов – 4 релевантных и 10 нерелевантных. Так как пользователь отыскал 2 нерелевантных документа из 10, то выпадение будет равно 20%.

В реальной жизни не существует систем, которые бы на практике обеспечивали идеальные показатели эффективности. Существующие системы либо не находят ВСЕ документы, либо находят нерелевантные документы, и пользователю приходится довольствоваться системой, способной в приемлемые сроки найти некоторое количество релевантных документов. Использование метода LSI позволяет улучшить качество поиска путем использования семантической модели информационных документов.

**Список литературы:** 1. *Ланкастер Ф.* Информационно-поисковые системы. М.: Мир, 1972. 308 с. 2. *Якушин Б.В.* Алгоритмическое индексирование в информационных системах. М.: Наука, 1978. 144 с. 3. *Новиков А.И., Ярославцева Е.И.* Семантические расстояния в языке и тексте М.: Наука, 1990. 320 с. 4. *Van Rijsbergen C. J.*. Information Retrieval. London: Butterworths, 1979. 386 p.

*Поступила в редколлегию 05.10.98*