Векторная модель представления текстовой информации

С. В. Моченов, А. М. Бледнов, Ю. А. Луговских
Ижевский государственный технический университет, Россия

The paper considers the approaches associated with the vector representation of the textual information. The particularity of the approach under consideration is in the determination of the goal functions of separate sentences and representation of them in the form of some local vectors on which basis a global vector is built that determines the semantic component of the text on the whole. Various aspects of application of the proposed approach are considered.

Введение

Широкое применение средств вычислительной техники в различных областях знаний сопровождается быстрым ростом объемов обрабатываемых массивов полнотекстовых документов и требует разработки новых подходов и средств организации доступа к информации.

Особую актуальность приобретает разработка методов извлечения и формирования новых знаний, необходимых для решения конкретных задач в той или иной профессиональной деятельности специалиста. Одним из стратегических направлений решения данной проблемы является комплексное системное использование различных лингвистических подходов и методов искусственного интеллекта, направленных на сокращение объемов хранимой информации, выявление семантической составляющей текста, определяющей основную идею, заложенную автором.

Основной задачей, возникающей при работе с полнотекстовыми базами данных, является задача поиска документов по их содержанию [Сокирко и др. 2005; Караулов и др. 1982; Финн 1999]. Существующие методы поиска, реализуемые, например, поисковыми машинами в Интернете, зачастую не обеспечивают адекватного выбора информации по запросу пользователя.

Пользователь не всегда может точно сформулировать поисковый запрос на получение информации, которая ему необходима. Более того, даже после получения этой информации требуется ее последующая аналитическая обработка с целью определения ее полезности и пригодности для решения поставленной задачи. Трудности, связанные с решением этой задачи, заключаются в многообразии возможных форм выражения одной и той же идеи, мысли, что особенно характерно для русскоязычных текстов.

В данной статье рассматриваются некоторые подходы к решению указанных проблем. Основное внимание уделено методам векторного представления текстовой информации.

1. Обзор методов векторного представления текстов

В конце 80-х годов в работах Салтона [Salton et al. 1994] была предложена векторная модель как альтернатива лексическому бесконтекстному индексированию. В простейшем случае векторная модель предполагает сопоставление каждому документу частотного спектра слов и соответственно вектора в лексическом пространстве. В процессе поиска частотный портрет запроса рассматривается как вектор в том же пространстве и по степени близости (расстоянию или углу между векторами) определяются наиболее релевантные документы.

В более продвинутых векторных моделях размерность пространства сокращается отбрасыванием наиболее распространенных или редко встречающихся слов, увеличивая тем самым процент значимости основных слов.

Главным достоинством векторной модели является возможность поиска и ранжирования документов по подобию, то есть по их близости в векторном пространстве. Однако практика показывает, что при оценке близости запроса к документу результаты поиска могут быть не всегда удовлетворительными, что особенно проявляется, когда запрос содержит малое количество слов. Для получения лучшей релевантности отклика в 1990 году была предложена модель скрытого семантического индексирования [Todd et al. 1999] — Latent Semantic Indexing (LSI). Модель использовала Singular Value Decomposition (SVD) для перехода от разреженной матрицы слов к компактной матрице главных собственных значений.

LSI показала значительное превосходство в результатах поиска по сравнению с лексическим методом, однако сложность модели часто приводила к существенному проигрышу в скорости на больших коллекциях документов по сравнению с традиционной булевой техникой [Salton 1989]. Одна из наиболее работоспособных систем на основе LSI была создана в Беркли в 1995 году Майклом Берри и Тодом Летче [Todd et al. 1995].

Описываемая ниже система использует совершенно другую интерпретацию понятия векторной модели текста, в которой не применяются частотные спектры слов.

2. Векторная модель представления текстовой информации

В данной работе текст рассматривается как структура, то есть как совокупность отдельных взаимосвязанных предложений, объединенных в подмножество абзацев, параграфов, глав и т. п. Эта структура обеспечивает выражение основной идеи, цели написания данного текста автором, через множество подцелей разного ранга, определяемых отдельными предложениями, абзацами, параграфами, главами и т. п.

Ниже приводятся примеры математической интерпретации векторной модели некоторого законченного элемента текста, состоящего, например, из глав, содержащих абзацы, которые, в свою очередь, состоят из предложений, то есть

$$G = \{G_1, G_2, \ldots, G_i, \ldots, G_n\}$$
$$Vg = \{Vg_1, Vg_2, \ldots, Vg_i, \ldots, Vg_n\},$$

где G — множество глав; G_i — *i-ая* глава, i = 1… n; Vg — множество векторов целей глав; Vg_i — вектор цели *i-ой* главы.

В свою очередь

$$A_i = \{A_{i1}, A_{i2}, \ldots, A_{ij}, \ldots, A_m\}$$
$$Va_i = \{Va_{i1}, Va_{i2}, \ldots, Va_{ij}, \ldots, Va_{im}\},$$

где A_i — множество абзацев *i-ой* главы; A_{ij} — *j-ый* абзац *i-ой* главы, j = 1 … m; Va_i — множество векторов целей абзацев; Va_{ij} — вектор цели *j-ый* абзаца *i-ой* главы.

Математическая интерпретация векторной модели для предложений выражается в виде:

$$P_{ij} = \{P_{ij1}, P_{ij2}, \ldots, P_{ijh}, \ldots, P_k\}$$
$$Vp_{ij} = \{Vp_{ij1}, Vp_{ij2}, \ldots, Vp_{ijh}, \ldots, Vp_{ijk}\},$$

где P_{ij} — множество предложений *i-ой* главы *j-го* абзаца; P_{ijh} — *h-ое* предложение *i-ой* главы *j-го* абзаца, h = 1 … k; Vp_{ij} — множество векторов целей предложений *i-ой* главы *j-го* абзаца; Vp_{ijh} — вектор цели *h-ого* предложения *i-ой* главы *j-го* абзаца.

Из представленного описания видно, что каждому элементу (фрагменту) текста ставится в соответствие некоторый вектор цели.

Как известно, смысловым и грамматическим центром предложения обычно является сказуемое, выраженное глаголом (полнозначным или связкой) [Бледнов и др. 2004; Арутюнова 2005]. При определенных условиях и именные группы (существительные с зависимыми словами или без них) могут выступать в качестве законченного предложения. Примерами таких предложений, называемых именными, или номинативными, являются, например:

Двадцать первое. Ночь. Понедельник. Очертанья столицы во мгле. (Ахматова).

Кроме именных предложений можно рассматривать и неполные, которые образуются из полных путем определенных сокращений. Например:

Отвертку! (вместо *Дай отвертку*!).

Причины такого сокращения, называемого *эллипсисом,* могут быть разнообразны, но обычно сокращается та часть предложения, которая рассчитана на определенные знания слушающего.

Как уже отмечалось выше, каждое предложение несет в себе определенный смысл, закладываемый автором, и обеспечивает продвижение к конечной цели, как основной идее в контексте системы целей смысловой группы предложений, абзаца и т. д. В общем случае каждое предложение имеет соответствующий вектор цели. Таким образом, текст можно определить как структуру взаимосвязанных понятий, обеспечивающую продвижение к конечной цели, выражаемой авторской идеей.

Предложенная модель вектора цели может быть представлена в виде трех компонент:

Vbegin — начальная цель, выражаемая через начальный вектор $\overline{X}$ с заданными координатами;

Vend — конечная цель, выражаемая через конечный вектор $\overline{Y}$ с заданными координатами;

Z — вид связи между начальным вектором $\overline{X}$ и конечным вектором $\overline{Y}$.

В качестве координат вектора могут выступать отдельные слова, понятия, именные группы, отдельные предложения, смысловые группы предложений, абзацы и т. д.

Поскольку имеются три составляющие, которые определяют вектор, то соответственно для последующего анализа нами были выделены следующие типы векторов:

1) простой вектор: $\overline{V} = (\overline{X})$ или V = ($\overline{Y}$);

2) нулевой вектор: $\overline{V} = (\emptyset)$;

3) полный вектор: $\overline{V} = (\overline{X}, \overline{Y})$ со связью Z;

4) пустой вектор: $\overline{V} = (\overline{X}, \overline{Y})$ без связи Z;

5) левый вектор: $\overline{V} = (\overline{X})$;

6) правый вектор: $\overline{V} = (\overline{Y})$.

В свою очередь векторы $\overline{X}$ и $\overline{Y}$ могут состоять из подвекторов, как отдельных самостоятельных частей — координат, принадлежащих текущему предложению.

Каждая координата имеет свои атрибуты *atr*. Атрибутами могут являться временные или пространственные характеристики координаты.

Состав координат вектора определяется сложностью построения предложения. В общем случае соподчиненность отдельных частей предложения может быть устранена путем нормализации.

Рассмотрим применение описанной выше векторной модели на конкретном примере.

Проанализируем следующее предложение.

Во все времена люди сталкиваются с одними и теми же проблемами экономики.

Данное предложение может быть представлено в векторной форме: $Vp(x_1; y_1)$ со связью вида z_1, или в упрощенной форме $Vp(x_1; y_1)(z_1)$,

где координата x_1 = {*люди*};

координата y_1 = {*проблемы экономики*};

связь вида z_1 = (*сталкиваются*).

При этом атрибутами координаты x_1 являются atr_x = (*во все времена*), а атрибутами координаты y_1 являются atr_y = (*одни и те же*).

На основе векторного представления могут быть решены некоторые проблемы обработки текстовой информации, в частности:

– сокращение объема исходной информации для выполнения процедур анализа текста и формирования систем и баз знаний;

– синтез текста с использованием информации, извлекаемой из баз знаний.

В следующем параграфе рассматривается геометрическая интерпретация технологии векторного представления текста.

3. Применение технологии векторного представления при анализе и синтезе текстовой информации

Рассмотренная выше векторная модель представления текстовой информации может быть использована при анализе и синтезе текста.

На рис. 1 представлена упрощенная интерпретация векторного представления текста в пространстве трех координат *x, y, atr*.

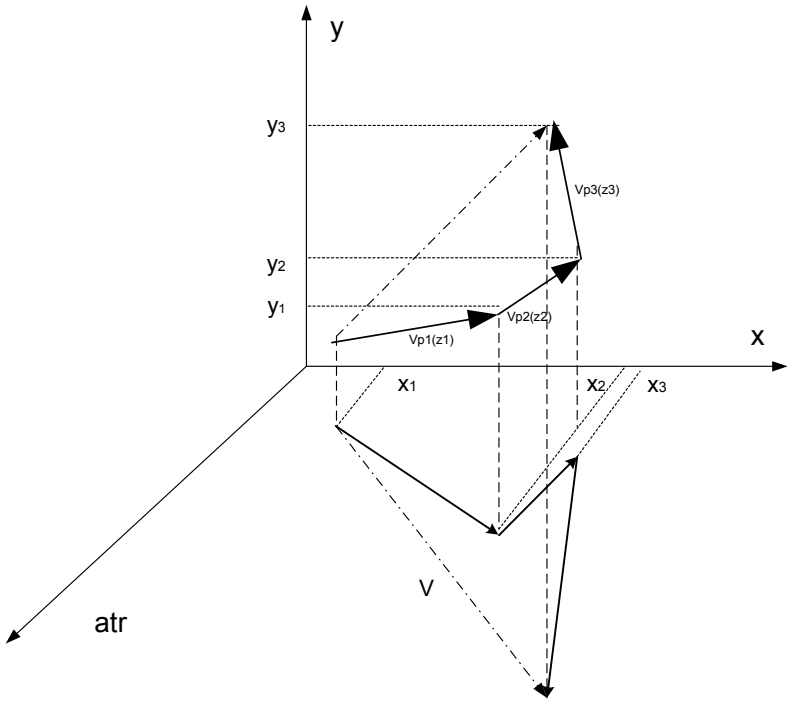


Рис. 1. Векторное представление текста

На представленном рисунке показаны три вектора $Vp_1(x_1, y_1)$ (z_1); $Vp_2(x_2, y_2)$ (z_2); $Vp_3(x_3, y_3)$ (z_3) и их проекция на плоскость *x, atr*. В общем случае проекцию можно осуществить на различные плоскости: (*x, y*), (*x, atr*), (y, atr).

Атрибуты могут иметь временные, пространственные и другие измеряемые характеристики.

Координаты x_i определяют начальные координаты вектора. Координаты y_i определяют конечные координаты вектора.

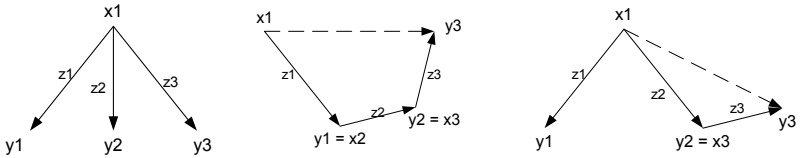
Исходя из предыдущего описания вектор V определяет конечную цель рассматриваемой единицы текста и имеет структуру вектора цели.

Рассмотрим использование описанной модели на примере.

Пусть заданы три вектора:

$Vp_1(x_1, y_1)$ (z_1); $Vp_2(x_1, y_2)$ (z_2); $Vp_3(x_1, y_3)$ (z_3).

На рис. 2 показаны некоторые возможные варианты геометрической интерпретации взаимодействия трех векторов.



а *б* *в*

Рис. 2. Геометрическая интерпретация взаимодействия трех векторов

Пример, представленный на рис. 2а, показывает, что координата x_1 вектора представляет собой иерархию понятий. Например, характеристика монитора: размер, количество цветов, производитель.

На рисунке 2б представлена геометрическая интерпретация взаимодействия другой группы из трех векторов:

$$Vp_1(x_1, y_1)\ (z_1);\ Vp_2(x_2, y_2)\ (z_2);\ Vp_3(x_3, y_3)\ (z_3),$$

где $y_1 = x_2$, $y_2 = x_3$.

Фактически совокупность этих трех векторов определяет некоторый результирующий вектор $V(x_1, y_3)$ (z'), соответствующий общей цели исходных векторов.

Следующий пример (рис. 2в) иллюстрирует независимость целей в приведенном наборе векторов:

$$Vp_1(x_1, y_1)\ (z_1);\ Vp_2(x_1, y_2)\ (z_2);\ Vp_3(x_3, y_3)\ (z_3),$$

где $y_2 = x_3$.

Другим применением векторной модели является возможность реализации синтеза текстовой информации.

Предположим, что решается задача, связанная с раскрытием понятия x_1. В этом случае, вектор цели для описания определенных процессов или явлений может быть представлен вектором $Vp_0(x_1, y_i)$ (z_i), где координата y_i и вид связей z_i, определяются в процессе конструирования подцелей.

Допустим, в базе знаний понятие x_1 определено на множестве онтологий через вектор $Vp_0(x_1, y_1)$ (z_1). В свою очередь подвектор y_1 имеет координаты $Vp_1(x_2, y_2)$ (z_2). Подвектор y_2 также может иметь свои координаты. Таким образом, получаем цепочку векторов для раскрытия понятия x_1. Механизм развертывания вектора для описания процессов и явлений может быть двояким: либо на основе сохраненного исходного текста, путем извлечения уже сформированных фраз, либо путем генерации новых предложений на основе алгоритма построения предложений на естественном языке.

На основе предложенной модели разработана технология обработки текстовой информации на основе векторной модели текста.

Заключение

Рассмотренные в данной статье основные положения технологии векторного представления текстовой информации и автоматизация этих процессов могут быть применены для:

– создания профессиональных систем и баз знаний;

– поддержки профессиональной деятельности работников различных отраслей;

– повышения уровня компетенции специалистов за счет получения возможности быстрого анализа и представления в удобной форме результатов этого анализа;

– проведения синтеза текстовых документов с различной степенью обобщения информации;

– автоматизации процессов формирования системы онтологий в той или иной профессиональной области;

– проведения направленного поиска и фильтрации текстовых документов;

– автоматического реферирования текстов документов.

Список литературы

Арутюнова 2005 — *Арутюнова, Н. Д.* Предложение и его смысл / Н. Д. Арутюнова. — М. : УРСС, 2005.

Моченов и др. 2005 — *Моченов, С. В.* Применение статистических методов для семантического анализа текста / С. В. Моченов, А. М. Бледнов, Ю. А. Луговских. — Ижевск : НИЦ «Регулярная и хаотическая динамика», 2005.

Караулов и др. 1982 — *Караулов, Ю. Н.* Русский семантический словарь. Опыт автоматического построения тезауруса: от понятия к слову / Ю. Н. Караулов, В. И. Молчанов, В. А. Афанасьев, Н. В. Михалев ; под ред. С. Г. Бархударова. — М. : Наука, 1982.

Рубашкин и др. 1998 — *Рубашкин, В. Ш.* Семантический (концептуальный) словарь для информационных технологий. Ч. 1 / В. Ш. Рубашкин, Д. Г. Лахути // НТИ. — Сер. 2. — 1998. — № 1. — С. 19–24.

Сокирко и др. 2005 — *Сокирко, А. Г.* Проект ДИАЛИНГ, COM-объект Goldrml / А. Г. Сокирко, Д. В. Панкратов. — М. : Диалог, 2005.

Финн 1999 — Финн В. К. О роли машинного обучения в интеллектуальных системах // НТИ. Сер. 2. 1999. № 12. — С. 1–3.

Salton 1989 — G. Salton. Automatic Text Processing. Addison-Wesley Publishing Company, Inc., Reading, MA, 1989.

Salton et al. 1994 — G. Salton, J. Allan, and C. Buckley. Automatic structuring and retrieval of large text files. Communications of the ACM, 37(2), February 1994.

Todd et al. — Todd A. Letsche and Michael W. Berry. Large-Scale Information Retrieval with Latent Semantic Indexing. URL: http://www.cs.utk.edu/ ~berry/sc95/sc95.html.

Проблемы описания и вопросы моделирования семантики слова в базах данных

И. М. Некипелова
Ижевский государственный технический университет, Россия

This article is devoted to the development in the field of modeling of the description of the word lexical value in the information retrieval system "Manuscript". This aspect is connected with the problem of use of the system for the implementation of linguistic research in the field of vocabulary and semantics and creation of the linguistic search system allowing the user making an exact idea about the word lexical value and its semantic relationships in the language and texts of the ancient manuscripts stored in a database.

В настоящее время актуальными в работе многофункциональных web-модулей транскрипций текстов являются разработки в области моделирования описания лексического значения слова и его семантических отношений. Творческая группа Удмуртского государственного университета и Ижевского государственного технического университета под руководством В. А. Баранова начала работы по созданию автоматизированного лексико-семантического модуля в информационно-поисковой системе «Манускрипт» (далее — ИПС «Манускрипт»). Это связано с необходимостью использования ИПС для проведения лингвистических исследований в области лексики и семантики и с необходимостью разработки лингвистической поисковой системы, позволяющей пользователю иметь точное представление о лексическом значении слова и его семантических связях в языке и текстах древних рукописей, хранящихся в базах данных ИПС «Манускрипт».

Особый интерес вызывают проблемы моделирования семантических и словообразовательных связей слов древних славянских и исходных древнегреческих текстов, поиск соответствий, хранение словообразовательных (морфемных и семантических) связей в базах данных и их использование. Практическое использование ана-