И.С. Лебедев

Построение семантически связанных информационных объектов текста

Развитие глобальных вычислительных сетей, сопряженное с необходимостью формирования больших объемов распределенных данных, делает весьма актуальной задачу автоматического анализа текстовой информации. В статье рассматриваются вопросы организации сети связанных объектов текстовой информации на основе использования семантического анализатора В.А. Тузова. Описываются правила построения, навигации и возможности использования таких конструкций для создания естественно-языковых интерфейсов в поисковых, справочных, обучающих системах. Значительное внимание уделяется рассмотрению методологических основ построения автоматических анализаторов текстовой информации. Автор отмечает, что при создании соответствующего программного обеспечения эффективным является комбинирование математических и лингвистических методов анализа. Изложение рассматриваемой методики сопровождается наглядными примерами и схемами, иллюстрирующими теоретические положения.

Автоматический анализ текстовой информации приобретает огромную актуальность в связи с развитием глобальных вычислительных сетей и формированием больших объемов распределенных данных.

Современные системы автоматической обработки текстов, доступные широкому кругу пользователей, например, информационно-поисковые машины в глобальных вычислительных сетях, в основном сталкиваются с проблемой классификации документов по запросу пользователя. На сегодняшний день существуют довольно приемлемые решения, дающие хорошие результаты, при анализе всего содержания документа в целом. Однако при разработке естественно-языкового интерфейса информационной системы подобные вещи необходимо решать по содержанию самого документа, вычисляя тот или иной абзац, множество предложений, где содержится ответ на вопрос пользователя. Исходя из этого разбиение текста на смысловые составляющие, определение семантических связей между ними является актуальной задачей.

Методы решения

Большинство решений данной задачи связано с использованием языков разметки, что требует от текста предварительной обработки экспертом, либо наличия жесткой структуры. Другие подходы к решению этой задачи заключаются в том, что текст представляется в виде информационного потока и по нему строится граф отношений, содержащий объекты текста и связи между ними.

Объекты текста, которые для простоты могут быть представлены словами, обозначаются соответствующими информационными элементами [1]. Одним и тем же словам соответствуют одинаковые информационные элементы. Простейшие системы, использующие подобные подходы, не содержат никаких словарей или тезаурусов, что позволяет достичь высокой скорости обработки за счет качества. На рис. 1 приведен пример текста и его графа.

Тестирование знаний путем проведения контрольных мероприятий является важным и необходимым элементом учебного процесса, однако в системе управ-

ления качеством результаты тестирования играют лишь вспомогательную роль. Действительно, тестирование непосредственно не указывает на причины и источники появления изъянов, оно является выборочным в отношении изучаемого материала и направлено преимущественно на оценку знаний и в меньшей мере на выявление умений обучаемых.

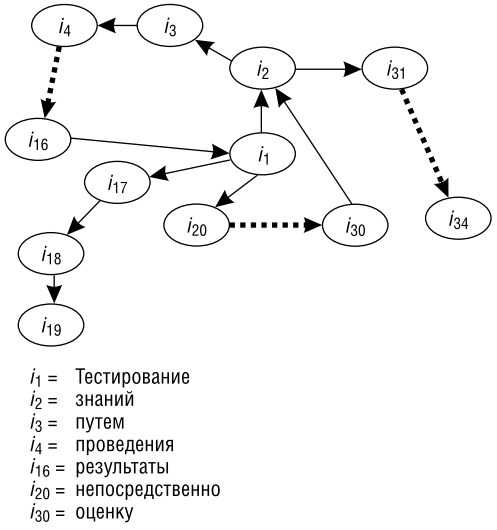


Рис. 1. Фрагмент графа текста[1]

В результате анализа, построенного по тексту графа, видно, что максимальное количество связей (по 4) образуют два информационных элемента «тестирование» и «знания», т. е. они являются основными для определения принадлежности тематики этого текста.

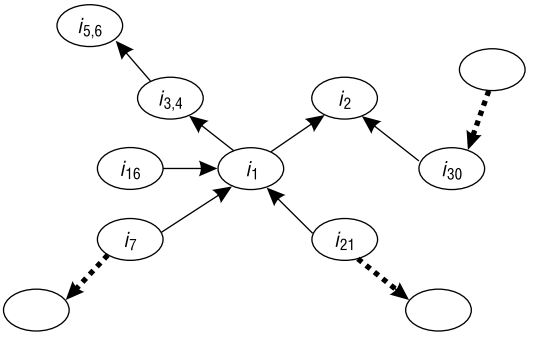
Однако даже такой простейший пример показывает, что для более точного определения связей необходимо проводить синтаксический и семантический анализ. Поэтому ниже приведен фрагмент обработки текста семантическим анализатором профессора СПбГУ В. А. Тузова, демонстрационная версия которого функционирует в сети [2].

```
является<X007.003>
  (@Им Тестирование
    (@Род знаний
      (@Род мероприятий
        (@Род контрольных
          (@КакВ путем
            (@Род проведения)
            ) ) ) ),
      @Тв элементом
      (@Тв важным
        @Тв и_необходимым,
        @Род учебного_процесса),
    однако_играют
    (@Им результаты
      (@Род тестирования),
    @Вин роль
      (@вПред в
        (@Пред системе_управления
          (@Тв качеством)
          ), @Вин лишь_вспомогательную)
        ) ) .
******************
не_указывает<X006.001>
  (@Им тестирование (Действительно),
  @КакВ непосредственно,
  @наВин на
  (@Вин причины,
    @Вин и_источники
    (@Род появления
    (@Род изъянов
      (@Род материала
        (@Род изучаемого
        (@ДееКак в_отношении)
          ) ) ) ) ),
является
  (@Им оно,
    @Тв выборочным),
      @Крат и_направлено
        (@КакВ преимущественно,
        @Куда на
        (@Вин оценку
        (@Род знаний) ),
        @ДееКак и_в
```

[1] В простейших системах анализа текста используется только порядок следования слов (в приведенном случае вершин графа) и основное внимание уделяется повторяющимся словам (узлам с максимальным количеством связей). Пунктирные стрелки обозначают последовательность следующих друг за другом вершин.

(*@Пред мере* (*@Пред меньшей*)
)),
@наВин на
(*@Вин выявление*
(*@Род умений*
(*@Род обучаемых*)
))).

Тогда граф текста примет вид, представленный на рис. 2.



i_1 = Тестирование
i_2 = знаний
$i_{3,4}$ = путем проведения
$i_{5,6}$ = контрольных мероприятий
i_7 = результаты
i_{21} = не указывает
i_{30} = оценку

Рис. 2. Граф текста после обработки анализатором[2]

При обработке данного текста семантическим анализатором становится очевидным, что «тестирование» образует 5 связей, а «знаний» — только 2. Семантический анализатор более точно распознает связи между словами.

Формализация описания модели

В формальном семантическом языке, в отличие от естественного, всякое слово рассматривается анализатором как некоторая функция f, значение которой определяется ее аргументами $x_1, \ldots, x_n$:

$$f(x_1, \ldots, x_n), \quad (1)$$

где $x_1, \ldots, x_n$ — слова, образующие конструкцию.

Например:

элемент (*какой*? — *необходимый*,
чего? — *учебного_процесса*)

Формализованное предложение — это конечный набор функций, связанный в единую суперпозицию [3]. Это означает, что предложение — также некоторая функция P, аргументами которой являются другие функции $f(x_1, \ldots, x_n)$, связанные между собой посредством определенных для них грамматических типов:

$$P(f_1(x_1, \ldots, x_n), f_2(x_1, \ldots, x_n), \ldots, f_3(x_1, \ldots, x_n)). \quad (2)$$

Грамматический тип, определяемый предлогом, предложной формой, позволяет определить синтаксические заголовки слов в формализованном словаре. Но эта информация будет неполной, если не определить формализованную роль частей речи в грамматической конструкции предложения. В формализованном синтаксисе все части речи (в отличие от естественного языка) равнозначны, нет главных и второстепенных членов предложения [4]. Аналогичным образом можно описать и семантические конструкции. Основное отличие будет состоять в том, что в семантическом словаре каждому слову приписывается свой идентификатор класса и жесткий набор классов, которые могут с ним употребляться и образовывать связи.

Каждый член предложения — функция со своими аргументами. Роль и поведение этих функций определяется значением их аргументов.

Таким образом, если $P(\{x_i\},\{y_j\})$, $i = 1 \ldots n$, $j = 1 \ldots k$, предложение, где $x_i \in X$ — множество слов, $y_j \in Y$ — множество конструкций слов, то обозначив через A множество описателей по предложению, необходимо найти такое их подмножество, которое конкретному набору слов предложения однозначно сопоставит его конструкцию.

Для решения этой задачи возможно использование абстрактно представленного

[2] Пустые вершины — это множества слов текста, образующие связи только между собой.

глагола как ведущей функции *G* в управлении предложением и морфологического описания всех возможных его аргументов.

$$G(\{10 \text{ аргументов основных падежей}\}, \text{где}, \text{зачем}, \text{как}, \text{какой}, \text{когда}, \text{который}, \text{куда}, \text{откуда}, \text{почему}, \text{сколько}, \text{чей}). \quad (3)$$

Аргументами абстрактной глагольной функции *G* выступает морфологическая информация о падежно-предложных формах слов в предложении.

$$P = G. \quad (4)$$

Таким образом задача анализатора сводится к нахождению адекватных морфологических и семантических описателей.

Алгоритм нахождения состоит из следующих этапов.

1. По каждому слову предложения, используя морфологический анализатор, находим его морфологический описатель m_k.

$$x_k \to m_k. \quad (5)$$

2. Если морфологические описатели совпадают, то этот набор и определяет соответствующую конструкцию *Y*, а это в свою очередь означает, что данный набор соответствует конкретному множеству описателей информации по данному предложению и, таким образом, достигается соответствие между словами и их конструкциями в предложении:

$$\{m_k\}_i : \{m_1 = m_2 = \ldots = m_i\} \to \{m_k\}_i \to \to Y_k \to \{m_k\}_i = A_k \to X_k \leftrightarrow Y_k. \quad (6)$$

Описание семантики синтаксиса предложений позволило разработать механизм сборки синтаксических конструкций в синтаксические шаблоны. Под синтаксическим шаблоном понимается такой способ представления информации о предложении в компьютере, по которой анализатор способен построить грамматически верную конструкцию предложения [5]. В дальнейшем синтаксические шаблоны послужили основой для словарных описателей семантического словаря, а механизмы сборки, адаптированные под семантическую модель, позволяют строить правильное дерево связей (граф) более 90% предложений естественного языка.

Описание структурных единиц текста

Наиболее сильно предложение характеризуют существительные. Они и представляют элементарные аргументы предложения. Их уточняют прилагательные, которые в свою очередь являются функциями со своими аргументами.

Существительные могут быть представлены в виде структуры, содержащей несколько полей.

$$S(k_1, \ldots, k_n). \quad (7)$$

где *S* — объект на основе существительного;

k — аргументы, которые присоединяются с помощью связей *какой*, *сколько*, *чей*, *чего*, *кого*, *кем*, *чем*.

Применительно к тексту, на котором проводится поиск, объекты можно условно классифицировать по нескольким типам.

1. Существительное, стоящее в тексте:

знания, *тестирование*

2. Существительное, уточненное прилагательным:

мероприятия контрольные

3. Существительные, уточненные другими существительными в родительном или творительном падеже:

результаты тестирования

4. Существительные с прилагательными, уточненные другими существительными в родительном или творительном падеже:

тестирование путем проведения мероприятий контрольных

Такое деление является относительным. Однако, если в запросе пользователя, заданного в любой форме, выделяются подобные группы на основе какого-либо су-

ществительного, то релевантный документ в своем тексте должен содержать слова уточняющей группы при этом существительном.

На основе тех форм запросов, которые выдают пользователи, применяя выражение (7) и подключив словарь синонимов, возможно, задавать перефразировки. Например, для запроса «результаты тестирования», используя электронный словарь синонимов [6], находим описания:

результат
следствие, *последствие*, *след*, *итог*, *плод*, *сумма*
тестирование
проверка, *испытание*

Подставив в выражение (7) получаем следующие перефразировки:

результаты тестирования, результаты проверки, результаты испытания, следствия тестирования, следствия проверки, следствия испытания, последствия тестирования, последствия проверки, последствия испытания, след тестирования, след проверки, след испытания, итог тестирования, итог проверки, итог испытания, плод тестирования, плод проверки, плод испытания, сумма тестирования, сумма проверки, сумма испытания…

Однако к подобным перефразировкам нужно относиться с осторожностью, так как в результате может возникнуть избыточность информации. Чтобы такого не происходило, словари синонимов должны подключаться только в соответствии с той тематикой, стилем и жанром, которые являются основными для текста. Кроме того, современный пользователь устроен таким образом, что он желает увидеть в ответе те же словоформы, что и в запросе, поэтому приоритет необходимо отдавать исходным словам.

Основой конструкции предложения, к которой прикрепляются все основные члены, является глагол. Если глагола в предложении нет, то его можно заменить глаголами типа «есть», взяв за основу «пустой глагол». Наречия уточняют глаголы. Такое описание позволяет рассматривать предложение, как глагольную функцию.

Каждый глагол аналогично существительным может быть также представлен в виде предиката:

$$N(G(S_1(k_1, \ldots, k_n), \ldots, S_m(k_1, \ldots, k_n))), \qquad (8)$$

где N — наречие, отвечающее на вопрос как, когда, куда, где, откуда, как долго;
G — глагольная функция;
S — объект на основе существительного.

Предложению, содержащему наречие, практически всегда возможно приписать один из шести вопросов.

Союзы в формализованном языке служат для установления связей между однородными словами в простом предложении и между простыми предложениями в составе сложного. Для анализатора важна информация о наличии какого-либо союза в произвольной конструкции и его семантическое назначение.

Для простого предложения достаточно формализации сочинительных союзов (информации о них). По значению они делятся на три разряда, описание которых указывает анализатору способ сборки конструкции и однородные аргументы конструкции.

Различают соединительные союзы (с.с.): они имеют значение соединения (и это, и то), и поэтому образуют составной аргумент по признаку его однородности, определяемый падежной формой. Тип такого аргумента формально выглядит так:

если $x_1 \neq x_2$, но $f(x_1) = f(x_2)$, то связка $f(x_1)$ [с.с.] $f(x_2)$ даст аргументную функцию

$$f_i(x_1) = f(x_1) + f(x_2). \qquad (9)$$

Ко второму разряду относятся противительные союзы (п.с.): они имеют значение противопоставления (не то, а это), поэтому:

если $x_1 \neq x_2$, но $f(x_1) = f(x_2)$, то связка $f(x_1)$ [п.с.] $f(x_2)$ даст аргументную функцию

$$f_i(x_1) = \begin{cases} f(x_1), \text{ если } f(x_2) = -f(x_2); \\ f(x_2), \text{ если } f(x_1) = -f(x_1). \end{cases} \qquad (10)$$

К третьему разряду относятся разделительные союзы (р.с.) (либо то, либо это). Их можно представить следующим образом:

если $x_1 \neq x_2$, но $f(x_1) = f(x_2)$, то связка $f(x_1)$ [р.с.] $f(x_2)$, также как и в первом случае, даст функцию

$$f_i(x_1) = f(x_1) + f(x_2). \quad (11)$$

Знаки препинания предназначены для уточнения и конкретизации аргументов. *Точка* для анализатора всегда является признаком окончания конструкции. Все, что находится за ней, считается новой конструкцией. *Двоеточие* служит анализатору признаком того, что в конструкции участвуют несколько одинаковых аргументов, по типу соответствующих первому, стоящему до двоеточия. *Запятая*, с одной стороны, служит признаком наличия в конструкции множества аргументов одного типа (в этом случае она аналогична союзу «и»), с другой — уточняет и разделяет сложные аргументы.

Применение модели на тексте

Текст представляет собой суперпозицию функций предложений. Современные алгоритмы и методы позволяют определить тематику целого текста, что приемлемо для систем классификации и рубрикации документов, но практически не применимо для проведения более глубокого анализа. Довольно часто конечного пользователя может интересовать из всего документа только одна строчка, которая подскажет необходимость ознакомления, например, «Описанные выше алгоритмы выделения памяти не применяются в операционных системах семейства *Windows*». Однако эту проблему можно решить созданием диалоговых систем.

При разработке вопросно-ответных систем, создании естественно-языкового интерфейса очень важно рассматривать не только документ в целом, а также его отдельные элементы. Причем для достаточно небольшого текста количество вхождений слов может быть мало информативно. В этом случае целесообразно использовать количество связей, которые они образуют. Так, в приведенном примере несколько слов (например, «тестирование», «знаний», «является») встречаются по два раза, а все остальные по одному, но слово «тестирование» образует сразу 5 связей. Подобный подход можно использовать при определении тех слов или терминов, на которые ложится основная смысловая нагрузка, что дает возможность выделить множество предложений, которые рассматриваются в качестве ответа на абстрактные вопросы по тексту, например: «Что сказано о тестировании?»

Если взять большой текст и построить график количества связей терминов (слов или словосочетаний) по предложениям или другим лексическим единицам текста, то можно увидеть, что анализируемый термин или слово встречается и образует связи неравномерно. На рис. 3 представлено количество связей для двух терминов текста некоторого технического описания объемом 15 предложений.

Рис. 3. Количество связей терминов по предложениям

Из рисунка видно, что в первом предложении интересующее нас словосочетание T_1 образовало три связи, во втором — две и т.д. Аналогично, можно сделать выборку для другого слова, например T_2. Теперь видно, в каких предложениях встретились термины, и где находится основная информация о них. Однако при ана-

лизе текстовой информации особенно на запрос, заданный в виде вопроса, есть вероятность, что ответ содержится в соседних предложениях, где искомый термин отсутствует. Для установки границ в рамках, которых просматривать предложения, можно вычислить порог p на основании отношения количества связей n к количеству предложений N.

$$p = \frac{n}{N}. \qquad (12)$$

Добавляя предложения, не содержащие термин, к множеству, где он присутствует, локализуем участок текста для дальнейшего анализа.

Будем считать, что если внутри множества предложений характеристики относительно определяемых объектов, меняются не существенно при добавлении новых предложений, то это множество является единым, для этих определяемых объектов.

Выбирая множество ближайших предложений, получаем некоторый портрет взаимосвязанных предложений, который во многих случаях является ответом на поставленный вопрос к тексту. Зачастую здесь содержится основная информация об объекте текста, которую можно узнать, привлекая только формально морфологические признаки.

Однако последнее время все чаще встречается мнение, что достичь качественного прорыва с применением одних только математических методов анализа текста не удается и все больше исследователей приходит к мнению о том, что необходимо подключать лингвистическую составляющую. Наиболее привлекательным, по причине простоты реализации, является деление текста на абзацы и предложения, но существуют другие трудоемкие, но более точные методы и их совокупности, например, определение кореференциальных связей [7], фиксирование последовательности развертывания текста, установление анафорических связей.

Если в запросе к тексту встретился только термин $S(k_1, \ldots, k_n)$ и $k = 0$, например «тестирование», то в качестве ответа приходится выдавать все предложения, где есть этот термин. Если в запросе $S(k_1)$ («тестирование знаний»), то для дальнейшего анализа выбираем предложения, содержащие $S(k_1)$, предложения, содержащие S («тестирование») и $S = k_1$ («знания») по отдельности, расположенные рядом друг с другом. На рис. 4 видна окрестность, которую образует термин «тестирование знаний» в тексте, приведенном в качестве примера в начале статьи.

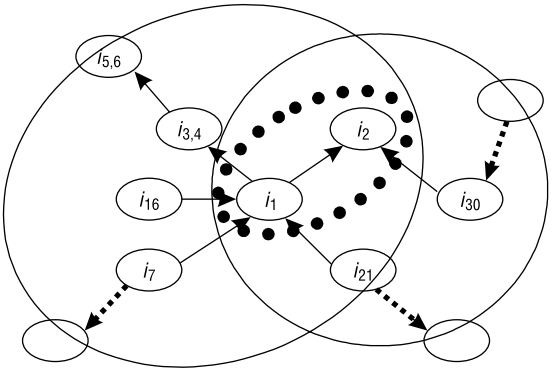


Рис. 4. Предложения, содержащие термины, и их окрестности

Внутри точечного овала находится термин «тестирование знаний», окружности ограничивают предложения, содержащие слова «тестирование» и «знания».

В тексте с большим количеством предложений может возникнуть ситуация, когда одно предложение будет содержать только термин S, а другое его аргумент существительное x_k. В этом случае необходимо анализировать являются эти предложения связанными между собой или нет.

Анализ отношений концентрируется вокруг нескольких основных вопросов: нахождение местоимений и других близких к ним слов заместителей; определение условий отождествления лексических повторов; учет актантной структуры предложения; отождествление имен [7].

При раскрытии какой-либо мысли в тексте ведется несколько параллельных рассуждений, которые в последствии вытекают в общие выводы. Предложения внутри небольшого фрагмента текста можно условно разделить на связанные, несвязанные и результирующие.

Анализ связей предложений текста можно производить на основе правил, изложенных ниже.

Предложения, связанные либо цепной, либо параллельной связью, определяются прямым перемещением основных слов, т. е. предыдущее предложение как бы содержит слово, которое стоит на ключевой позиции в данном предложении, или повторяет его основу. В безличных предложениях основное ударение переводится либо на существительное в винительном падеже, либо, в случае отсутствия, на существительное родительного или дательного падежа. Именительный падеж однозначно указывает основной объект в предложении, относительно которого строится предложение. При переходе смыслового ударения обычно в одном предложении указывается новый объект, а в другом идет повторение нового смыслового слова в основе предложения. Если идет передача смысла через глагол, то связка осуществляется через слова-определители: для этого, тогда, пусть и т. д.

Несвязанные предложения не имеют общих слов.

Результирующие предложения обычно начинаются со слов «потому», «поэтому», «для этого», «чтобы» и т. д. Они содержат обобщающую мысль абзаца текста. Могут стоять и перед предложениями, подводящими к результату.

Выводы

Связь предложений в тексте, в случае ее формализации дает возможность определить границы текста, где можно анализировать несколько предложений в качестве ответа на вопрос.

Для анализа текста в вопросно-ответных системах необходимо получить как можно более полный и точный граф предложений.

Анализ графа предложений можно комбинировать математическими и лингвистическими методами.

Используя граф предложений и правила его анализа, становится возможным частично отражать текст в базу знаний или базу данных и автоматически строить связи.

Для более точного отражения текста в базу знаний или базу данных необходимо проводить более точный анализ связей внутри набора предложений, выделенных предлагаемыми методами.

Список литературы

1. *Чугреев В.Л.*, *Яковлев С.А.* Анализ текста, применительно к решению задач поиска документов по образцу // *Информатизация процессов формирования открытых систем на основе САПР, АСНИ, СУБД и систем искусственного интеллекта (ИНФОС-2003): Материалы II Международной научно-технической конференции.* Вологда: ВоГТУ, 2003.

2. Проект SemLP. Демонстрация системы; http://www.semlp.com

3. *Тузов В.А.* Компьютерная семантика русского языка. СПб.: Изд-во СПбГУ, 2004.

4. *Кондратьев А.В.*, *Кривцов А.Н.*, *Лебедев И.С.* Анализаторы текстов формальной модели русского языка для компьютера // *Процессы управления и устойчивость: Труды XXIX научной конференции студентов и аспирантов факультета ПМ-ПУ.* СПб.: НИИ Химии СПбГУ, 1998.

5. *Комаров И.И.*, *Кривцов А.Н.*, *Лебедев И.С.* Принципы построения семантической модели текста и ее применение в системах лингвистического обеспечения // *Процессы управления и устойчивость: Труды XXXIII научной конференции студентов и аспирантов факультета ПМ-ПУ.* СПб.: НИИ Химии СПбГУ, 2002.

6. Информационный сервер г. Набережные Челны. Электронный словарь синонимов; http://www.chelni.ru/slovari/sinonim

7. *Рубашкин В.Ш.* Представление и анализ смысла в интеллектуальных информационных системах. М., 1989.