УДК 519. 816

Н.К. Тимофеева

НЕКОТОРЫЕ СПОСОБЫ МОДЕЛИРОВАНИЯ ВХОДНЫХ ДАННЫХ ДЛЯ ПОИСКА ИНФОРМАЦИИ В БИБЛИОТЕКЕ ЭТАЛОНОВ ПРИ РЕШЕНИИ ЗАДАЧ СЕМАНТИКИ

Ключевые слова: задачи по семантике, комбинаторная оптимизация, кластеризация, покрытие заданными признаками определенного объекта, поиск эталона в базе даних.

Введение

Рассматриваются задачи семантики, которые относятся к задачам распознавания и для установления сути определенного объекта требуют нахождения его эталона в базе данных. Показано, что моделирование этих задач как задач комбинаторной оптимизации позволяет достаточно строго описать предметную область и показать, что поиск подходящего эталона в базе данных проводится двумя способами: по первичным признакам, которые описывают искомый объект, и по заданному объекту. Также описаны задачи, в которых входные данные разделяются на сегменты с последующим определением сходства полученных частей. Входная информация в этом случае содержит и объект, который необходимо распознать, и эталон, с которым он сравнивается. Решение задачи проводится без библиотеки эталонов.

Постановка проблемы и цель исследования

Для установления сути объекта при решении задач семантики необходимо сравнить его с эталоном, который, как правило, находится в библиотеке эталонов. Возникает две задачи. Разработка способов поиска эталона в реальном времени и нахождение эффективных подходов для установления сути объекта без библиотеки эталонов.

Задача поиска эталона относится к *NP*-полным задачам. Поэтому возникает проблема сведения поставленной задачи к разрешимой путем структуризации библиотеки по определенным правилам. Есть задачи, в которых эталонную информацию можно задать аналитически или алгоритмически. Тогда для сравнения достаточно смоделировать входную информацию по тем же правилам, что и информация, которая играет роль эталона. Такой подход позволяет полностью автоматизировать процесс решения поставленной задачи.

Анализ последних исследований и публикаций

Теории информационного поиска посвящено огромное количество литературы, например, [1, 2]. При моделировании информационных потоков изучаются структурные связи между массивами документов, которые в них входят. В настоящее время для моделирования этих связей используют фрактальный анализ, методы нелинейной динамики, теории клеточных автоматов, методы самоорганизации. В теории информационного поиска рассматриваются также проблема семантики см. например, [3–5]. Как правило, к этому направлению относят задачи, связанные с языкознанием. Эта проблема рассматривается в компьютерных науках, искусственном интеллекте. В вычислительном интеллекте анализируются задачи, которые требуют установления сути предмета, но способы их решения не всегда позволяют достичь поставленной цели.

Для установления сути предмета используют эталонную информацию, которая содержится в базах данных и представляет собой структурированную совокупность взаимосвязанных данных определенной предметной области. Для быстрого нахождения в ней необходимой информации эта база должна быть соответственно структурирована. Представление информации о предметной области связано с моделированием данных. Существуют различные модели данных, которые имеют свои преимущества и недостатки, и каждая из них имеет свою область применения.

В теории моделирования данных при решении задач семантики применяется метод, названный семантическим моделированием. Он заключается в моделировании структуры данных, опираясь на их содержание, что важно для интеллектуализации различных систем и позволяет глубже вникать в сущность определенного объекта.

В качестве инструмента семантического моделирования используются различные варианты диаграмм сущность–связь или ER-диаграммы [3, 5]. Для правильного применения ER-диаграмм создаются математические модели, формулировка которых базируется на основе таких математических понятий, как теория множеств, теория решеток, теория графов. Тип сущности интерпретируется как множество, а сущность — как элемент этого множества.

При структуризации и поиске информации в базе данных возникают задачи покрытия определенными признаками объектов и кластеризации. Если детально проанализировать сравнение входной информации и эталона, то можно увидеть, что поиск соответствующего эталона в базе данных проводится двумя способами: по первичным признакам, которые описывают искомый объект, и по заданному объекту. Также существуют задачи семантики, решение которых не требует библиотеки эталонов.

Задачи семантики

1. Распознавание речи — это процесс автоматической обработки речевого сигнала для определения последовательности слов, которая передается этим сигналом. Задача заключается в нахождении для входного сигнала наиболее правдоподобного эталона из всех возможных эталонных сигналов [6]. Путем сравнения устанавливается дословное их сходство без анализа смыслового значения слов (предложения).

2. Распознавание детского, женского, мужского голосов. Задача распознавания детского, женского, мужского голосов решается путем анализа сигнала на значение амплитуды, длины периода основного тона. Эта задача разрешима, поскольку оговоренные параметры можно описать достаточно строго и задать их по условию. В этой задаче устанавливается суть предмета, поэтому ее можно отнести к задачам семантики.

3. Многодикторное распознавание речи. Речевые сигналы, соответствующие одному и тому же слову, но произнесенные разными дикторами, отличаются как частотой, так и величиной амплитуды [6]. В этой задаче в процесе распознавания сравниваются входной сигнал и эталонный. Но для распознавания необходимо проводить адаптацию эталонов к голосу нового диктора. Эта задача частично относится к семантическому анализу, поскольку необходимо распознать индивидуальный голос.

4. Задача клинической диагностики. Эта задача состоит в нахождении для множества признаков, характеризующих заболевания пациента, наиболее правдоподобного одного или нескольких эталонов из множества заболеваний, т.е. по входным признакам устанавливается одно или несколько заболеваний [7]. Поскольку в этой задаче устанавливается суть объекта, отнесем ее к задачам семантики.

5. Сравнение текстов с целью установления плагиата. Существующие программные средства сравнивают одинаковые слова или фразы. Если текст совпадает с оригиналом, то программы достаточно просто выявляют плагиат. Если суть (значение) текста, который анализируется, остается такой же, как и в оригинале, но передана другими фразами (словами), то при обнаружении плагиата необходимо проводить семантический анализ обоих текстов. Эта задача относится к семантике.

6. Криптография, дешифрование забытых письменностей и т.д. относится к задачам семантики. На начальном этапе дешифровки объект воспринимается без определения его сути. На втором этапе (распознавание) отдельно воспринимаются и анализируются составляющие признаки объекта и определяется суть данного объекта. Третий этап (интерпретация) — заключительный, наиболее сложный этап дешифровки, во время которого анализируются и обобщаются количественные и качественные признаки. Смысловая сторона дешифровки не всегда поддается автоматизации.

7. Автоматический перевод текстов с одного языка на другой. Перевод текстов проводится двумя способами: дословный и художественный. Первый подход не является задачей семантики, так как в результате получаем дословный перевод без анализа его сути и он поддается автоматизации. Но, как правило, перевод, осуществленный таким образом, некачественный. Во втором подходе для осуществления качественного перевода необходимо проводить анализ текста на суть предмета. Но в этом случае автоматизировать художественный перевод текстов достаточно сложно.

Некоторые из рассмотренных задач сводятся к задачам комбинаторной оптимизации. Для них построим математические модели в рамках этой теории. Аргументом целевой функции у них являются комбинаторные конфигурации разных типов.

Общая математическая постановка задачи комбинаторной оптимизации

Задачи комбинаторной оптимизации, как правило, задаются на одном или нескольких множествах, например $A = \{a_1, ..., a_n\}$ и $B = \{b_1, ..., b_{\tilde{n}}\}$, элементы которых имеют любую природу [8]. Назовем эти множества базовыми. Имеется два типа задач.

1. Каждое из этих множеств представим в виде графа, вершинами которого являются его элементы, а каждому ребру поставлено в соответствие число $c_{lt} \in R$, которое называют весом ребра (R — множество действительных чисел); $l \in \{1, ..., n\}$, $t \in \{1, ..., \tilde{n}\}$, n — количество элементов множества A, $\tilde{n}$ — количество элементов множества B. Положим, что $n = \tilde{n}$. Между элементами этих множеств существуют связи, числовое значение которых назовем весами. Величины $c_{lt} \in R$ — входные данные, зададим их матрицами.

2. Между элементами заданного множества связей не существует, а весами являются числа $v_j \in R$, $j \in \{1, ..., n\}$, которым соответствуют некоторые свойства этих элементов, числовые значения которых задаются конечными последовательностями и являются входными данными.

Для обоих типов задач из элементов одного или нескольких базовых множеств, например $a_l \in A$, $l \in \{1, ..., n\}$, образуется комбинаторное множество W — совокупность комбинаторных конфигураций определенного типа (перестановки, выборки различных типов, разбиения и т.д.). На элементах w комбинаторного множества W вводится целевая функция $F(w)$. Необходимо найти элемент w^* множества W, для которого $F(w)$ принимает экстремальное значение при выполнении заданных ограничений.

Под комбинаторной конфигурацией понимаем любую совокупность элементов, которая образуется из всех или некоторых элементов заданного базового множества $A = \{a_1, ..., a_n\}$ [8]. Обозначим ее упорядоченным множеством $w^k = (w_1^k, ..., w_\eta^k)$, где $\eta \in \{1, ..., n\}$ — количество элементов в w^k, $W = \{w^k\}_1^q$ — множество комбинаторных конфигураций. Верхний индекс k $(k \in \{1, ..., q\})$ в w^k обозначает порядковый номер w^k в W, q — количество w^k в W.

Смоделируем структуру входных данных функциями натурального аргумента. Представим элементы h наддиагоналей симметричной комбинаторной матрицы $Q(w^k)$ комбинаторной функцией $\beta(f(j), w^k)|_1^m = (\beta_1(f(1), w^k), ..., \beta_m(f(m), w^k))$, а элементы h наддиагоналей симметричной матрицы C — функцией натурального аргумента $\varphi(j)|_1^m = (\varphi(1), ..., \varphi(m))$, где $m = \frac{n(n-1)}{2}$ — количество элементов h наддиагоналей матриц C и $Q(w^k)$, $h = \overline{1, n-1}$. Если матрицы $Q(w^k)$ и C — несимметричные, то $\beta(f(j), w^k)|_1^m$ и $\varphi(j)|_1^m$ содержат все их элементы, а $m = n^2$ (или $m = n\tilde{n}$). Функция цели примет вид $F(w^k) = \sum_{j=1}^{m} \beta_j(f(j), w^k)\,\phi(j)$

Математическая модель поиска эталона в базе данных по признакам, которые задаются как входные данные

При семантическом моделировании возникает задача максимального покрытия объекта определенными признаками, которые его характеризуют [9]. Она решается для дальнейшего установления сути этого объекта путем поиска определенного эталона в базе данных. Признаки разделяются на такие, которые характеризуют лишь заданный объект, по которым достаточно просто его определить в базе данных. В этом случае задача является разрешимой. Если одинаковые признаки описывают различные объекты, но с помощью дифференциального анализа можно найти искомый объект, то такая задача является частично разрешимой. Если одни и те же признаки характеризуют различные объекты и по ним нельзя идентифицировать искомый, то возникает ситуация неопределенности.

Рассмотрим подробнее задачу покрытия соответствующими признаками определенного объекта. Смоделировав ее в рамках теории комбинаторной оптимизации можно увидеть, что она относится к задачам разбиения, аргументом целевой функции в которой является разбиение n-елементного множества A на подмножества как с повторениями так и без повторений. Ее можно смоделировать и как задачу кластеризации.

Пусть задана база данных с объектами (эталонами) различной природы. Обозначим их множеством A. Также заданы признаки, характеризующие эти объекты. Обозначим их множеством B. Выделим следующие подзадачи.

- Объекты из множества A покрываются признаками из B полностью или частично так, чтобы последние не пересекались.
- Объекты из множества A покрываются признаками из B так, чтобы последние полностью покрывали заданные объекты. В этом случае может возникнуть ситуация, когда один и тот же элемент из B может характеризовать различные объекты.

В первой задаче образованные кластеры не пересекаются, т.е. $w_p \cap w_l = \varnothing$. Во второй образованные кластеры могут пересекаться, т.е. $w_p \cap w_l \neq \varnothing$.

Задача покрытия заключается в отыскании такого разбиения $w^* \in W$, при котором заданный объект максимально покрывается минимальным количеством признаков, по которым можно установить его сущность при выполнении условия, а именно: количество одинаковых в разных кластерах элементов минимально.

Построим математическую модель задачи поиска эталона в базе данных по признакам, которые их характеризуют, с использованием теории комбинаторной оптимизации. Обозначим $A = \{a_1, ..., a_n\}$ множество объектов, описание которых находится в базе данных (множество эталонов), где элемент $a_t \in A$, $t \in \{1, ..., n\}$, соответствует определенному объекту, которому соответствуют его характерные признаки $V^{(t)} = (v_1^{(t)}, v_2^{(t)}, ..., v_{q'_t}^{(t)})$, q'_t — количество признаков *t*-го объекта. Входной информацией в этой задаче является множество признаков $\tilde{V} = (\tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_{\tilde{q}})$, которые могут описывать один или несколько различных объектов. Обозначим их $B = \{b_1, ..., b_{n'}\}$, где $b_p \in B$ — объект, который нужно определить, n' — количество возможных объектов, а $q'_t \geq \tilde{q}$. Признаки $\tilde{v}_r \in \tilde{V}$ входной информации имеют тот же смысл, что и эталонные: $v_l^{(t)} \in V^{(t)}$, $r \in \{1, ..., \tilde{q}\}$, $l \in \{1, ..., q'_t\}$, $p \in \{1, ..., n'\}$. Возможна ситуация, когда в библиотеке эталонов отсутствуют объекты, признаки которых поступают со входной инфомацией.

Задача заключается в нахождении для B со множеством признаков $\tilde{V}$ наиболее правдоподобного одного или нескольких эталонов из множества $A = \{a_1, ..., a_n\}$, т.е. по входных признакам устанавливается один или несколько объектов $b_p \in B$. Признаки в этой задаче играют роль критериев, по которым оценивается ее решение. Необходимо провести поиск определенного эталона в библиотеке и сравнить его со входными признаками. В этом случае основная задача разделяется на две подзадачи.

Рассмотрим случай, когда входная информация касается одного объекта. Для него смоделируем целевую функцию. Обозначим выражением $u_l(v_s^{(t)}, \tilde{v}_r)$ элементарную меру сходства между элементами множеств $\tilde{V}$ и $V^{(t)}$. Поскольку эти признаки могут иметь различные шкалы измерений, выберем такие, чтобы полученные оценки сводились к одной шкале. Полагаем, что меры сходства $u_l(v_s^{(t)}, \tilde{v}_r)$ принимают значения $\{0, ..., 1\}$, $r \in \{1, ..., \tilde{q}\}$, $s \in \{1, ..., q'_t\}$. Если для однотипных элементов $v_s^{(t)}$ и $\tilde{v}_r$ $u_l(v_s^{(t)}, \tilde{v}_r) \in \{\delta, ..., 1\}$, будем считать, что они одинаковые, где δ — наименьшая величина степени сходства, для которой существует допустимое решение. Однотипные элементы $v_s^{(t)}, \tilde{v}_r$, для которых $u_l(v_s^{(t)}, \tilde{v}_r) < \delta$, будем считать разными. Если множества $\tilde{V} \bigcap V^{(t)} = \varnothing$, то они не содержат никаких одинаковых элементов $v_s^{(t)}$ и $\tilde{v}_r$. Если $\tilde{V} \bigcap V^{(t)} \neq \varnothing$, то множества $\tilde{V}$ и $V^{(t)}$ содержат одинаковые элементы. Их может быть один и больше.

Рассмотрим задачу сравнения признаков $V^{(t)} = (v_1^{(t)}, v_2^{(t)}, ..., v_{q'_t}^{(t)})$, описывающих *t*-й эталон, и входных признаков $\tilde{V} = (\tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_{\tilde{q}})$, по которым устанавливается определенный объект. Будем считать, что меры сходства между элементами $v_s^{(t)} \in V^{(t)}$ и $\tilde{v}_r \in \tilde{V}$ являются входными данными. Их числовые значения зададим конечной последовательностью (комбинаторной функцией натурального аргумен-

та), зависящей от размещения без повторений w^k. Обозначим ее $\beta(f(j), w^k)|_1^{n^*} = (\beta_1(f(1), w^k), ..., \beta_{n^*}(f(n^*), w^k))$, где $n^* = \min(q_t', \tilde{q})$. Если элементы $v_s^{(t)} \in V^{(t)}$ и $\tilde{v}_r \in \tilde{V}$ однотипные, то значение $\beta_j(f(j), w^k) = u_l(v_s^{(t)}, \tilde{v}_r)$. В противном случае $\beta_j(f(j), w^k) = 0$.

Для определения подобных элементов из множеств $V^{(t)}$ и $\tilde{V}$ введем комбинаторную функцию $\tilde{\beta}(\varphi(j), w^k)|_1^{n^*} = (\tilde{\beta}_1(\varphi(1), w^k), ..., \tilde{\beta}_{n^*}(\varphi(n^*), w^k))$, $\varphi(j) \in \{0, 1\}$, где $\tilde{\beta}_j(\varphi(j), w^k) = 1$, если $\beta_j(f(j), w^k) \geq \delta$. В противном случае $\tilde{\beta}_j(\varphi(j), w^k) = 0$. Количество единиц в функции $\tilde{\beta}(\varphi(j), w^k)|_1^{n^*}$ равно $q''(w^k) = \sum_{j=1}^{n^*} \tilde{\beta}_j(\varphi(j), w^k)$.

Оценку результата проводим по одному критерию, для которого запишем следующие целевые функции: среднее значение мер сходства, если $q''(w^k) > 0$:

$$F_1(w^k) = \left(\sum_{j=1}^{n^*} \tilde{\beta}_j(\varphi(j), w^k) \beta_j(f(j), w^k) \right) \Big/ q''(w^k), \quad (1)$$

суммарное значение мер сходства:

$$F_2(w^k) = \sum_{j=1}^{n^*} \tilde{\beta}_j(\varphi(j), w^k) \beta_j(f(j), w^k). \quad (2)$$

Здесь $\sum_{j=1}^{n^*} \tilde{\beta}_j(\varphi(j), w^k) \beta_j(f(j), w^k)$ — интегральная мера сходства, а $\beta(f(j), w^k) = u_l(v_s^{(t)}, \tilde{v}_r)$ — элементарная мера сходства, которая определяет сходство между элементами эталона и элементами множества признаков входных данных. Аргументом целевых функций (1), (2) является размещение без повторений.

Множество размещений без повторений W состоит из подмножеств изоморфных комбинаторных конфигураций W_η. Задача сравнения входных признаков с эталонными производится на всем множестве W. Как оговорено в [8], закономерность изменения значений целевых функций (1), (2) на заданном упорядочении W подмножествами W_η одинакова, независимо от входных данных. В этом случае в процессе решения задачи возникает ситуация неопределенности, связанная с моделированием целевой функции и структурой ее аргумента. Если по функциям (1), (2) получаем оптимальное решение для одной и той же $w^{k*} \in W_\eta \subset W$, то результат совпадает с целью исследования. В противном случае использование функций (1), (2) приводит к ситуации неопределенности.

Задача сравнения эталона и входных признаков заключается в поиске такого размещения без повторений $w^{k*} = (w_1^{k*}, ..., w_{\eta^{k*}}^{k*})$, для которого смоделированные функции (1), (2) принимают наибольшие значения.

В задаче перебора эталонов $A = \{a_1, ..., a_n\}$ как веса между элементами $a_t \in A$ и входными данными $\tilde{V}$ выступают величины интегральных мер сходства, полученные по выражениям (1), (2), числовое значение которых представим

матрицами. Номера столбцов этих матриц совпадают с номерами эталонов $a_t \in A$, размещенных в библиотеке. Строка в них одна и соответствует номеру один множества $V = \{\tilde{V}\}$. Поскольку при сравнении множества признаков входных данных и признаков эталонов из базовых множеств A и V выбирается два элемента независимо от их упорядочения, образованная комбинаторная конфигурация является сочетанием без повторений. Задача поиска библиотечного эталона, соответствующего входным признакам, состоит в отыскании такого сочетания без повторений, для которого значение смоделированной целевой функции, по которой оценивается результат решения, было бы наибольшим.

Как видно из постановки задачи, поиск эталона, подобного исходному $\tilde{V}$, требует полного перебора. Эту задачу можно свести к разрешимой путем структуризации библиотеки эталонов по определенным признакам, которые определяют предметную область, т.е. на этапе структурирования библиотеки решается задача кластеризации, аргументом целевой функции в которой является разбиение *n*-элементного множества на подмножества.

Для задачи кластеризации сформулируем такие условия. Рассмотрим множество подмножеств $\rho = (\rho_1, ..., \rho_\eta)$ такое, что $\rho_1 \cup ... \cup \rho_\eta = A$, $\rho_p \cap \rho_l = \varnothing$, $p \neq l$, $\rho_p \neq \varnothing$, $p, l \in \{1, ..., \eta\}$, $\eta \in \{1, ..., n\}$. Непустое подмножество $\rho_p = \{a_1, ..., a_{\xi_p}\}$, $a_s \in A$, $s \in \{1, ..., n\}$, может иметь от 1 до n элементов. Количество подмножеств ρ_p в разбиении ρ также может быть от 1 до *n*.

Задача кластеризации заключается в разбиении элементов заданного множества A на кластеры так, чтобы смоделированная целевая функция принимала оптимальное значение. Оптимизация производится по критерию наибольшего сходства между признаками, которыми покрывается определенный объект.

Изложенная математическая модель описывает задачу клинической диагностики.

Математическая модель поиска эталона в базе данных, который соответствует заданному объекту без предварительного покрытия определенными признаками

Поставленная задача не требует предварительного покрытия признаками объекта и эталона. Входные и эталонные данные содержат признаки по своей природной сути, по которым устанавливается их сходство. Например, распознавание речевых сигналов производится без предварительного покрытия их признаками. Рассмотрим эту задачу.

Как было оговорено, распознавание речи — это процесс автоматической обработки речевого сигнала для определения последовательности слов, которая передается этим сигналом [6]. Опишем его последовательностью $X = (x_1, ..., x_n)$, элемент x_i которой является значением сигнала в отсчете *i*. Длина *n* различных реализаций сигнала определенного слова — разная. Для распознавания из различных реализаций X создается словарь эталонных слов. Эталон слова словаря описывается последовательностью $E_h = (e_{h_1}, ..., e_{\hat{q}_h})$, где h — номер слова в словаре, $\hat{q}_h$ — длина сигнала эталона слова, $h \in \{1, ..., \ddot{q}\}$, $\ddot{q}$ — количество эталонных сигналов в библиотеке.

Задача распознавания речевых сигналов заключается в отыскании для сигнала X наиболее правдоподобного эталона E_h из всех возможных эталонных сигналов. Задача распознавания речевых сигналов естественно разделяется на две

подзадачи: перебор эталонных сигналов и сравнение эталонного и входного сигналов. Поскольку здесь имеет место перебор вариантов, то она относится к задачам комбинаторной оптимизации.

Рассмотрим задачу сравнения эталонного и входного сигналов. Введем два базовых множества: $A = \{a_1, ..., a_n\}$ и $B = \{b_1, ..., b_{\tilde{n}}\}$, где $a_i = x_i \in X$, $i = \overline{1, n}$, а $b_l = e_{h_l} \in E_h$, $l \in \{1, ..., \tilde{n}\}$. Положим, что $n = \min(n, \tilde{n})$, а $m = n^2$. Входными данными для установления сходства оговоренных сигналов являются веса между элементами $a_i \in A$ и $b_l \in B$. Зададим их несимметричной матрицей $C = \| c_{il} \|_{n \times n}$, номера столбцов которой совпадают с нумерацией элементов $a_i \in A$, а номера строк — с нумерацией элементов $b_l \in B$. Поскольку из каждого базового множества A и B выбирается по одному элементу в строгом порядке, то полученная комбинаторная конфигурация — размещение без повторений. Обозначим ее $\mu^k \in M$, где M — их множество. Для определения элементов $a_i \in A$ и $b_l \in B$, которые выбираются из базовых множеств на k-м варианте решения задачи, введем комбинаторную (0, 1) матрицу $Q(\mu^k) = \| g_{il}^k(\mu^k) \|_{n \times n}$. Если $g_{il}^k(\mu^k) = 1$, то из множеств A и B выбрана пара (a_i, b_l), иначе — значение $g_{il}^k(\mu^k) = 0$. Для записи целевой функции в явном виде смоделируем входные данные функцией натурального аргумента. Элементы матрицы C представим числовой функцией $\varphi(j)|_1^m$, а матрицы $Q(\mu^k)$ — комбинаторной $\beta(f(j), \mu^k)|_1^m$.

Задача сравнения эталонного и входного сигналов заключается в отыскании такого размещения без повторений $\mu^{k*} = (\mu_1^{k*}, ..., \mu_{q'}^{k*})$, для которого функционал

$$F(\mu^{k*}) = \max_{\mu^k \in M} \sum_{j=1}^{m} \varphi(j) \beta_j(f(j), \mu^k), \tag{3}$$

где $\sum_{j=1}^{m} \varphi(j) \beta_j(f(j), \mu^k)$ — интегральная мера сходства, а $\varphi(j) = g'_j(a_i, b_l)$ — элементарная мера сходства, которая определяет сходство между элементами эталонного и входного сигналов.

Рассмотрим задачу поиска эталонного сигнала, подобного входному.

Обозначим A и $\tilde{B} = \{B_1, ..., B_{\ddot{q}}\}$, базовые множества, где $A = X$, а $B_l = E_{h_l}$. В этой задаче как веса между эталонным и входным сигналами выступают значения интегральных мер сходства, полученных по выражению (3), числовое значение которых представим матрицей C'. Номера столбцов этой матрицы совпадают с номерами эталонных сигналов, размещенных в библиотеке. Строка в ней один и соответствует номеру один входного сигнала. Поскольку при сравнении входного и эталонного сигналов из базовых множеств A и B выбирается два элемента, то образованный объект — сочетание без повторений. Обозначим его $\mu'^k \in M'$, где M' — их множество. Введем комбинаторную (0, 1)-матрицу $Q(\mu'^k) = \| g_{1l}'^k(\mu'^k) \|_{1 x \ddot{q}}$. Если $g_{1l}'^k(\mu'^k) = 1$, то из множеств A и B выбрана пара (A, B_l), иначе — значение

$g'^{k}_{1l}(\mu'^{k}) = 0$. Элементы матрицы C' представим числовой функцией $\varphi'(j)|_1^{n-1}$, а матрицы $Q(\mu'^{k})$ — комбинаторной $\beta'(f'(j), \mu'^{k})|_1^{n-1}$.

Задача поиска эталонного сигнала, который соответствует входному, заключается в нахождении такого сочетания без повторений $\mu'^{t*} = (A_1, B_l)$, для которого значение заданной целевой функции было бы наибольшим, т.е.

$$F(\mu'^{t*}) = \max_{\mu'^{k} \in M} \sum_{j=1}^{n-1} \varphi'(j)\beta'_j(f'(j), \mu'^{t}), \tag{4}$$

где $\varphi'(j) = \sum_{j=1}^{n^*} \varphi(j)\beta_j(f(j), \mu^k)$.

Многодикторное распознавание речи. Речевые сигналы, соответствующие одному и тому же слову, но произнесенные разными дикторами, отличаются как частотой, так и величиной амплитуды [6]. Для распознавания необходима адаптация к голосу нового диктора. Эта задача частично относится к семантическому анализу, поскольку необходимо распознать индивидуальный голос. Представим эту задачу с использованием мультимножеств.

Речевое пространство рассмотрим как свернутое, содержащее базовое множество A (активные и пассивные органы создания речи) и правила, по которым комбинацией элементов множества A возникают речевые сигналы (развернутое речевое пространство) [10]. Точками речевого пространства является выборка — размещение с повторениями из n элементов $a_s \in A$ по η, в котором учитывается порядок элементов, $s, \eta \in \{1, ..., n\}$. Одно и то же слово, повторенное несколько раз одним и тем же диктором или различными дикторами, отличается благодаря тому, что полученное размещение с повторениями содержит разное количество элементов. Отсюда — нечеткость во входных данных. Итак, речевой сигнал, который является входной информацией, описывается комбинаторной конфигурацией (размещение с повторениями). Представим его мультимножеством. Оно формально определяется как пара (A, m) где $m: A \to N$ — функция из A в множество N натуральных чисел, т.е. каждому элементу множества A соответствует определенное натуральное число, которое называется кратностью этого элемента.

Речевой сигнал зададим последовательностью $f|_1^{\tilde{n}} = (f_1, f_2, ..., f_{\tilde{n}})$, где f_i — значение амплитуды в отсчете i сигнала. Проведем его сегментацию на почти периодические и непериодические отрезки. Текущий почти период разделим на k отсчетов и опишем мультимножеством, которое зададим основой $(f|_1^k, m)$, где k — величина, определяемая экспериментально и должна быть одинаковая для любого отрезка сигнала. В *i*-м отсчете должно быть только одно значение f_i. Эталон, по которому устанавливается сходство почти периода, моделируется аналогично. По выражениям $|f_i - f'_i| \le \varepsilon$ и $|m_i - m'_i| \le \varepsilon'$ устанавливаем сходство входного сигнала и эталона, где f'_i — значение сигнала эталона в отсчете i, m'_i — кратность элемента f'_i, ε, ε' — минимальные величины, при которых может быть установлено подобие сигналов. Они определяеются экспериментально.

Приведем вычислительную схему решения этой задачи.

• Проведем сегментацию входного сигнала на почти периодические и непериодические участки алгоритмом [11].

• Текущий почти период разделим на k отсчетов и опишем его мультимножеством, которое зададим основой $(f\,|_1^k, m)$.

• Эталон, по которому устанавливается сходство почти периода, моделируется аналогично.

• Определим сходство входного и эталонного сигналов по выражениям $\left| f_i - f_i' \right| \le \varepsilon$ и $\left| m_i - m_i' \right| \le \varepsilon'$.

Эта задача относится к семантике и поддается автоматизированному распознаванию голоса разных дикторов. Сравнение входного и эталонного сигналов можно проводить методами, описанными в [6].

Итак, процесс поиска эталона в библиотеке для обеих задач, который соответствует заданному объекту, проводится без предварительного их покрытия определенными признаками.

Определение сути объекта без библиотеки эталонов

Сегментация речевого сигнала. Особенность этой задачи заключается в том, что сигнал разделяется на сегменты таким образом, чтобы последние были подобными. В этом случае входная информация содержит объект, который необходимо распознать, и эталон, с которым устанавливается сходство. Таким образом, суть объекта определяется условием, по которому вводятся коэффициенты сходства, определяющиеся экспериментально. Для решения задачи сегментации речевого сигнала разработано много методов и алгоритмов, основанных на корреляционных подходах с использованием динамического программирования, например [6]. Но во многих подходах она решается и путем распознавания конфигурации входного сигнала. Проведем сегментацию речевого сигнала на почти периодические и непериодические участки, где распознается его конфигурация [11]. Эта задача заключается в выделении на заданном отрезке входного сигнала почти периодических и непериодических участков, а в почти периодических определяются длины текущего почти периода. Для формулировки математической постановки этой задачи используем теорию комбинаторной оптимизации.

Отрезок исследуемого сигнала разобьем на участки длиной $L \in \{L_{\min}, L_{\min} + \Delta, L_{\min} + 2\Delta, ..., L_{\max}\}$ с последующим определением почти периодичности соседних участков; $L_{\min}$ — минимально возможная длина почти периода, $L_{\max}$ — максимально возможная длина почти периода, Δ — значение прироста почти периода (определяется экспериментально). Интегральная степень сходства, которая используется для определения почти периодичности двух соседних участков, формулируется с учетом нескольких критериев.

Пусть задан отрезок речевого сигнала L, который представим в виде числовой функции $f(j)\,|_1^m$, m — количество ее значений. Будем полагать, что соседние участки (объект и эталон) функции $f(j)\,|_1^m$ почти периодические, если меры сходства $p_l = \dfrac{\min(a_t - a_{t-1}, a_{t-1} - a_{t-2})}{\max(a_t - a_{t-1}, a_{t-1} - a_{t-2})} > \varepsilon$, $d_l = \dfrac{\min(b_t - b_{t-1}, b_{t-1} - b_{t-2})}{\max(b_t - b_{t-1}, b_{t-1} - b_{t-2})} > \varepsilon$, $\pi_l = \left| p_l - d_l \right| < \varepsilon'$, где $t \in \{3, ..., m(\mu^k) + 1\}$, $l \in \{1, ..., m(\mu^k)\}$, p_l, d_l — элементарные меры сходства, которые позволяют определять почти периодичность t-го и $(t+1)$-го участков; a_t — отсчет, для которого значение функции $f(a_t)$ на участке длиной $\tilde{L}$ наибольшее, или $\left|\, \left| f(a_t) \right| - \left| f(x^*) \right| \,\right| < \varepsilon''$, если $\dfrac{\min(a_t - a_{t-1}, \tilde{L})}{\max(a_t - a_{t-1}, \tilde{L})} > \varepsilon$,

x^* — отсчет, для которого значение функции $f(j)$ наибольшее; b_t — отсчет, для которого значение функции $f(b_t)$ на этом же участке наименьшее или $\big|\big|f(b_t)\big| - \big|f(b^*)\big|\big| < \varepsilon''$, если $\dfrac{\min(b_t - b_{t-1}, \tilde{L})}{\max(b_t - b_{t-1}, \tilde{L})} > \varepsilon$, b^* — отсчет, для которого значение функции $f(b^*)$ — наименьшее; $\varepsilon, \varepsilon', \varepsilon''$ — коэффициенты меры сходства, определяемые экспериментально; $m(\mu^k)$ — количество участков длиной $\tilde{L}$, на которые разбивается функция $f(j)|_1^m$ на k-й итерации. В процессе решения задачи формируется размещение без повторений $\mu^k = (\mu_1^k, ..., \mu_{\eta^k}^k)$ (аргумент целевой функции), в которой учитывается порядок элементов, причем $\mu_\sigma^k < \mu_{\sigma+1}^k$, $\eta^k \in \{1, ..., m(\mu^k)\}$ — количество элементов в μ^k. Элемент $\mu_\sigma^k \in \{1, ..., m\}$ — отсчет сигнала, который указывает на конец σ-го или на начало $(\sigma+1)$-го почти периода $\sigma \in \{1, ..., \eta^k - 1\}$. Оценка и выбор оптимального варианта решения задачи из всех возможных μ^k проводится по целевой функции, в которой учитывается несколько критериев.

Распознавание детского, женского, мужского голосов. Задача распознавания детского, женского, мужского голосов производится путем анализа сигнала на значение амплитуды и длины периода основного тона. Эта задача разрешима, поскольку оговоренные параметры можно описать достаточно строго. В ней устанавливается суть предмета, поэтому ее можно отнести к задачам семантики. В этой задаче в некоторых случаях может быть ситуация неопределенности, если мужчина имеет женский голос, и наоборот, — женщина имеет мужской голос. Представим вычислительную схему ее решения.

1. Проведем сегментацию входного сигнала, который описывается комбинаторной конфигурацией (размещение с повторениями), на периодические и непериодические участки алгоритмом [11], а в периодических выделим почти периоды.

2. По результатам сегментации установим среднюю длину периода основного тона (частоты сигнала).

3. Определим среднее значение амплитуды сигнала.

4. Сравним полученные данные с коэффициентами, заданными по условию.

5. Установим тип голоса (детский, женский или мужской).

Описанная задача является разрешимой и не требует для своего решения библиотеки эталонов.

Заключение

Итак, поиск эталона в библиотеке эталонов для решения задач семантики проводится двумя способами: по первичным признакам, которые описывают искомый объект, и по заданному объекту. Эти задачи сводятся к задачам комбинаторной оптимизации. При описании искомого объекта по первичным признакам возникают такие задачи, как покрытие и кластеризация. Моделирование этих задач, как задач комбинаторной оптимизации, позволяет достаточно строго описать предметную область. Существуют задачи по распознаванию, которые могут быть задачами семантики, входные данные которых разделяются на сегменты с последующим определением сходства полученных частей. Входная информация в этом случае содержит и объект, который необходимо распознать, и эталон, с которым он сравнивается. В некоторых задачах за эталон принимают либо выражение, по

которому определяют сходство входной и эталонной информации, либо задают условия, по которым можно распознать заданный объект. Решение таких задач проводится без библиотеки эталонов.

Н.К. Тимофієва

ДЕЯКІ СПОСОБИ МОДЕЛЮВАННЯ ВХІДНИХ ДАНИХ ДЛЯ ПОШУКУ ІНФОРМАЦІЇ В БІБЛІОТЕЦІ ЕТАЛОНІВ ПРИ РОЗВ'ЯЗАННІ ЗАДАЧ СЕМАНТИКИ

Для встановлення суті предмета використовують еталонну інформацію, що міститься в базах даних та є структурованою сукупністю взаємопов'язаних даних певної предметної області. Для швидкого знаходження в ній необхідної інформації ця база має бути структурована, а також відповідно змодельовані і вхідні дані. Нині існують різні моделі даних зі своїми перевагами та недоліками, і кожна з них має свою область застосування. Наведено приклади задач семантики, що відносяться до задач розпізнавання і для встановлення суті певного об'єкта потребують знаходження його еталона в базі даних. Це — розпізнавання мовлення, розпізнавання дитячого, жіночого, чоловічого голосів, задача клінічної діагностики, порівняння текстів на плагіат, автоматичний переклад текстів з однієї мови на іншу тощо. Порівняння вхідної інформації та еталона проводиться двома способами: за первинними ознаками, які описують шуканий об'єкт, і за заданим об'єктом. При другому способі попереднє покриття певними ознаками еталона і об'єкта не проводиться. При моделюванні вхідних даних для пошуку інформації за першим способом має місце покриття певними ознаками заданих об'єктів. Ознаки розділяються на такі, які характеризують лише заданий об'єкт, за якими досить просто його визначити в базі даних. В цьому разі задача є розв'язною. Якщо однакові ознаки описують різні об'єкти, але за допомогою диференціального аналізу можна знайти потрібний об'єкт, то така задача є частково розв'язною. Якщо одні і ті ж ознаки характеризують різні об'єкти і за ними не можна ідентифікувати шуканий, то виникає ситуація невизначеності. Існують задачі, які для свого розпізнавання не потребують бібліотеки еталонів. У деяких задачах з розпізнавання вхідні дані розділяються на сегменти з подальшим визначенням подібності отриманих частин. Вхідна інформація в цьому разі містить і об'єкт, який необхідно розпізнати, і еталон, з яким він порівнюється. У деяких задачах за еталон приймають або вираз, за яким визначають подібність вхідної та еталонної інформації, або задають умови, за якими можна розпізнати заданий об'єкт. Для розв'язання таких задач бібліотека еталонів не використовується.

Ключові слова: задачі семантики, комбінаторна оптимізація, кластеризація, покриття заданими ознаками певного об'єкта, пошук еталону в базі даних.

N.K. Tymofijeva

SOME WAYS TO MODELING INPUT DATA FOR INFORMATION SEARCH IN A MODEL LIBRARY WHEN SOLVING SEMANTICS PROBLEMS

To determine the essence of the subject it is used standard information contained in databases which is a structured set of interconnected data of a specific subject area. To find quickly the information it needs, this database should be structured, input data should also be modeled accordingly. Today there are input data models that have

their advantages and disadvantages, and each model has its own scope. The article provides examples of problems of semantics that relate to recognition problems. To determine the essence of a particular object, it requires finding its standard in the database. These are speech recognition, child, female, male voice recognition, the problem of clinical diagnostics, comparison of texts on plagiarism, automatic translation of texts from one language to another, etc. There are two ways of comparing the input information and the standard: by the primary signs that describe the object being sought and by the given object. In the second method previous cover by certain signs of the standard and of the object is not conducted. At modeling input data to search for information by the first way the cover of given objects by certain signs takes place. The signs are divided into those that characterize only the given object, by which it is quite simply to define it in the database. In this case the problem is solvable. If the same signs describe different objects, but using differential analysis you can find the the desired object, then this problem is partially solvable. If the same signs characterize different objects and the desired object cannot be identified, then a situation of uncertainty arises. There are problems that do not require a standard library for their recognition. In some recognition problems, which may be semantics problems, the input data is divided into segments, with subsequent determination of the similarity of the resulting parts. In this case, the input data contains both the object to be recognized and the standard with which it is compared. In some problems it is taken as a standard either an expression that determines the similarity of the input data and standard information, or the conditions are specified by which a given object can be recognized. The standard library is not used to solve these problems.

Keywords: semantics problems, combinatorial optimization, clustering, coating by a given signs of a particular object, database search used.

1. Ланде Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. М. : Либроком (Editorial URSS), 2009. 264 с. ISBN 978-5-397-00497-8:
2. Manning C., Raghavan P., Schütze H. Introduction to information retrieval. Cambridge University Press, 2008. 544 p. ISBN 0-521-86571-9.
3. Дейт К. Дж. Введення в системи баз даних. 8-е видання. М. : Видавничий дім «Вільямс», 2005 1328 с:
4. Кудрявцев Д.В. Системы управления знаниями и применение онтологий. Санкт-Петерб.: Политехн. ун-т, 2010. 340 с.
5. Сільвейструк Л.М. Формалізація моделі «сутність-зв'язок: типи сутностей, типи зв'язків та їх обмеження. Автореф. дис ... канд. фіз-мат. наук. К., 2009. – 18 с.
6. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. Киев: Наук. думка, 1987. 262 с.
7. Тимофієва Н.К., Гриценко В.И. Аргумент цільової функції в задачі клінічної діагностики. *Управляющие системы и машины*. 2012. № 3. С. 3–14.
8. Тимофієва Н.К. Теоретико-числові методи розв'язання задач комбінаторної оптимізації. Автореф. дис. ... докт. техн. наук. Київ, 2007. 32 с.
9. Тимофієва Н.К. Формалізація семантичного моделювання з використанням теорії комбінаторної оптимізації. Інформаційні технології та комп'ютерне моделювання (ІТКМ–2018). *Матеріали міжнародної науково-практичної конференції.* Івано-Франківськ–Яремче, 14-19 травня 2018 р. Івано-Франківськ. 2018. С. 6–9.
10. Тимофієва Н.К. Знакові комбінаторні простори та штучний інтелект. *Штучний інтелект.* 2015. 67-68, (1-2). С. 180–189.
11. Тимофієва Н. Ітераційний алгоритм автоматичного визначення квазіперіодичних і неперіодичних ділянок мовного сигналу. *Оброблення сигналів і зображень та розпізнавання образів.* Третя Всеукр. Міжнар. конф. Київ, 26–30 листопада 1996. К., 1996. С. 132–134.

Получено 26.06.2020