# 2. РАЗРАБОТКА МОДЕЛИ СТРУКТУРНОГО ПРЕДСТАВЛЕНИЯ И МЕТОДА ТЕМАТИЧЕСКОГО АНАЛИЗА ТЕКСТА

Задача поиска документов по образцу предпологает решение двух основных задач:

1) тематическая классификация текстовой информации;

2) вычисление степени тематической принадлежности текста к заданному классу.

Эти задачи связаны, прежде всего, с анализом текста, а именно, с анализом смыслового содержания текста, его тематической направленности.

Всю совокупность представленных на сегодняшний день методов анализа текста, относительно задачи анализа его содержания, можно разделить на две большие группы:

* лингвистический анализ;
* статистический анализ.

Первый ориентирован на извлечении смысла текста по его семантической структуре. Второй – по частотному распределению слов в тексте.

В данной работе было принято решение использовать методы статистического анализа в силу их относительной простоты, удобства использования и языковой независимости. Методы лингвистического анализа, хотя и позволяют точнее анализировать текст, выделяя его структурные особенности, но являются более трудоемкими и сложными в использовании. Связано это, прежде всего, с богатством семантики и морфологии естественных языков. Формальное описание правил естественного языка и их реализация – весьма трудоемкий процесс, требующий привлечения специалистов из области лингвистики. Кроме того, лингвистический анализ предполагает ориентацию на конкретный язык с его конкретными семантическими особенностями, это обуславливает его плохую межъязыковую переносимость. Работы в данном направлении идут [10, 24, 123], и существует множество практических реализаций, но на сегодняшний день лингвистический анализ по части анализа семантики весьма проблематичен.

Все это обусловило целесообразность применения статистических методов для решения задач данной работы. Однако частотный анализ, используемый в настоящее время при определении тематики документов [120], не позволяет в полной мере учесть внутреннюю структуру текста, т.к. при таком анализе не учитывается связность и последовательность текста. Хотя именно связность текста (речевого высказывания) считается одним из важнейших условий, необходимых для понимания его смыла и содержания. Данное положение является ключевым как в психолингвистике [3, 27], так и нейропсихологии [1, 18].

В ряде работ [11, 28] в области информационного поиска также отмечается эта особенность. Опираясь на результаты этих исследований, а также результаты собственных исследований [29, 30, 32], автором была разработана модель структурного представления текста, учитывающая его связность [34, 35].

## 2.1. Графовая модель структурного представления текста произвольного содержания

Суть предлагаемого подхода заключается в моделировании структуры текста информационным потоком и формировании этим потоком ориентированного мультиграфа, вершинами которого являются слова, а ребрами – связи между словами в тексте. Этот мультиграф является информационной структурой текста [34, 35].

Мультиграф – это граф, который может содержать множество ребер соединяющих одну и туже пару вершин.

Информационный поток – это детерминированный поток событий, принадлежащих некоторому конечному множеству. Временной интервал между событиями нас не интересует, интересует только последовательность событий. В данном случае события – это слова, а конечное множество – это множество всех слов, присутствующих в анализируемом тексте. Информационный поток эквивалентен временному ряду номинальных (категориальных) величин.

Под информационной структурой понимается совокупность всех событий и связей между ними. Для информационной структуры текста связи между событиями – это словосочетания.

Информационный поток, по сути, моделирует динамику некоторого процесса, в данном случае текста, а информационная структура является статическим представлением информационного потока.

Переход к модели структурного представления текста осуществляется следующим образом.

1) Текст рассматривается в виде информационного потока, образованного информационными элементами - словами.

Если последовательно брать слова из текста, начиная с самого первого и кончая последним, то это как раз и будет информационный поток F.

При этом набор всех слов в тексте можно выделить в конечное множество уникальных информационных элементов:

I = {i1, i2, …, in},

где i – информационный элемент соответствующий определенному слову из текста.

Информационный поток F будет представлен в виде последовательного чередования этих элементов (рис. 2.1).

rn-1

r2

r1

Рис. 2.1. Информационный поток

F = (i1, i2, …, in).

Информационный поток также может быть представлен в виде набора связей:

F = (r1, r2, …, rn-1),

где: ri = (ii, ii+1) – связь между двумя информационными элементами, последовательно идущими в информационном потоке.

Далее, чтобы не перегружать рисунки избыточной информацией, будем опускать именование связей.

Порядок чередования информационных элементов зависит от их последовательности в тексте. Информационные элементы в потоке могут повторяться. Обязательное условие – однозначное соответствие информационного элемента слову из текста. Одинаковые слова в тексте соответствуют одному и тому же информационному элементу.

Пример.

Фрагмент текста: “Дао, которое может быть выражено словами не есть постоянное Дао. Имя, которое может быть названо, не есть постоянное имя”.

Из данного набора слов выделяем множество уникальных информационных элементов (различия в регистре и знаки препинания не учитываются): I = {i1, i2, i3, i4, i5, i6, i7, i8, i9, i10, i11},

где:

i1 = быть,

i2 = выражено,

i3 = дао,

i4 = есть,

i5 = имя,

i6 = которое,

i7 = может,

i8 = названо,

i9 = не,

i10 = постоянное,

i11 = словами.

Информационный поток соответствующий данному фрагменту текста (рис. 2.2) будет представлен в виде набора информационных элементов:

F = (i3, i6, i7, i1, i2, i11, i9, i4, i10, i3, i5, i6, i7, i1, i8, i9, i4, i10, i5)

Рис. 2.2. Информационный поток, соответствующий заданному тексту

2) Поток формирует структуру.

Если учесть, что слова в тексте повторяются, то, соответственно, можно допустить, что информационный поток будет многократно проходить через одни и те же информационные элементы, формируя, таким образом, связанную информационную структуру текста.

Для вышеприведенного фрагмента текста данная структура будет выглядеть следующим образом (рис. 2.3).

Рис. 2.3. Структура, формируемая информационным потоком

Тезис о том, что информационный поток будет многократно проходить через одни и те же информационные элементы, носит принципиальный характер, при отсутствии циклов графа, образованных информационным потоком, ни о какой формируемой структуре не приходится говорить.

На самом деле структура, представленная на рис.2.3, не отражает в полной мере информационный поток, описывающий фрагмент текста. Необходимо дополнить эту структуру информацией о прохождении потока через каждый из информационных элементов. Для каждого повторного прохождения потока через одну и ту же пару информационных элементов необходимо формировать дополнительные связи – ребра. Тогда структура описывается в виде мультиграфа – графа, который может содержать множество ребер соединяющих одну и ту же пару вершин.

Для удобства отображения такого мультиграфа проиндексируем информационный поток и припишем каждому ребру графа, соединяющего пару вершин, множество индексов, соответствующих прохождению информационного потока через данную пару.

Рассмотрим в качестве примера все тот же фрагмент текста, который описывается информационным потоком:

F = (i3, i6, i7, i1, i2, i11, i9, i4, i10, i3, i5, i6, i7, i1, i8, i9, i4, i10, i5).

Индексация данного информационного потока будет означать, что каждому переходу между двумя информационными элементами будет поставлен в соответствие индекс, начиная с единицы, с последовательным его инкрементом.

5

17

8

16

7

13

3

12

2

15

11

4

14

6

18

9

10

1

Рис. 2.4. Индексированная структура информационного потока

Информационная структура, приведенная на рис. 2.4, в полной мере отражает все особенности потока, описывающего фрагмент текста.

Весьма показательным является приведенный выше фрагмент текста, особенность его, прежде всего, в том, что для сравнительно небольшого количества слов, входящих в данный фрагмент, существуют циклы. И информационный поток, моделирующий данный фрагмент, формирует явно выраженную структуру. Лингвистический анализ текста, его стилистические особенности и смысловая многозначность выходят за рамки диссертационной работы, но некоторая корреляция, по-видимому, существует между информативностью и структурной насыщенностью - наличием большого числа связей между отдельными элементами. Явно выражен принцип усложнения системы за счет введения дополнительных связей между отдельными элементами.

Введем дополнительные обозначения и определим некоторые важные характеристики информационной структуры.

n(I) = | I | - количество информационных элементов множества I (количество уникальных слов в тексте).

n(F) = | F | - количество информационных элементов набора F (общее количество слов в тексте).

M(I, R) – информационная структура (ориентированный  мультиграф). Является совокупностью I - множества информационных элементов (вершин графа) и R - набора связей между этими элементами (ребер графа).

M(I, R)  F.

R - набор связей между парами информационных элементов, может содержать повторяющиеся связи в случае многократного прохождения информационного потока F через одни и те же пары элементов.

R = (r1, r2, …, rn-1),

где: r = (ii, ii+1) – связь между двумя информационными элементами, обозначает последовательность информационных элементов ii, ii+1 в потоке F.

.

Для каждого информационного элемента из множества I, входящего в структуру M(I, R), существует набор пар связей, где: ri, rj – входная связь, ri+1, rj+1– выходная связь, n – число пар связей.

Входная означает, что данная связь предшествует выходной в наборе связей, описывающих поток, проходящий через данный информационный элемент. Если проиндексировать связи в наборе, описывающие поток, то индекс входной связи будет на единицу меньше выходной.

n(R(i)) - количество пар связей в наборе R(i).

Характеризует число связей данного информационного элемента с другими информационными элементами в структуре M(I, R). n(R(i)) равно числу повторений слова в тексте.

Обозначим количество пар связей как:

d(i) – степень информационного элемента

d(i) = n(R(i)),

, .

d(M(I, R))max – максимальная степень информационного элемента для информационной структуры M(I, R):

d(M(I, R))max = max d(i), .

d(M(I, R))min – минимальная степень информационного элемента для информационной структуры M(I, R):

d(M(I, R))min = min d(i), .

r+(F, ik, ij, e), r-(F, ik, ij, e) – расстояние между двумя информационными элементами в потоке F, , где:

e – вхождение информационного элемента ik в поток F, его порядковый номер в потоке.

r+ - означает, что расстояние измеряется по ходу информационного потока.

r- - означает, что расстояние измеряется обратно ходу информационного потока.

Расстояние измеряется в количестве информационных элементов находящихся между ik и ij,плюс 1.

Пример:

F = (i2, i5, i1, i4, i5, i2, i3)

r+(F, i2, i3, 1) = 6

r+(F, i2, i3, 2) = 1

r-(F, i1, i2, 1) = 2

r-(F, i1, i3, 1) = 0

r+(F, i1, i3, 1) = 4

0 – означает, что результат не определен

rmin(F, ik, ij) – минимальное расстояние между двумя информационными элементами в потоке, .

Пример:

F = (i2, i6, i1, i4, i5, i7, i2)

rmin(F, i2, i5) = 2

rmin(F, i1, i2) = 2

rmin(F, i1, i8) = 0

Информационный поток относительно некоторого информационного элемента i можно описать как:

F(i, e, [r-, r+]),

где e – вхождение информационного элемента ik в поток F, его порядковый номер в потоке; [r-, r+] – окрестность, для которой определяется поток.

F(i, e, [r-, r+]) = (i - r-, …, i - 2, i - 1, i, i + 1, i + 2, …, i + r+),

где  - обозначает индексацию некоторого информационного элемента в наборе F относительно информационного элемента i;

i + 1 – обозначает информационный элемент, следующий сразу за i в информационном потоке F;

i - 1 – обозначает информационный элемент, предшествующий i, в информационном потоке F.

Пример:

F = (i2, i5, i1, i4, i5, i2, i3, i8, i10, i1)

F(i4, 1, [3, 3]) = (i2, i5, i1, i4, i5, i2, i3)

F(i2, 2, [2, 4]) = (i4, i5, i2, i3, i8, i10, i1)

F(i2, 1, [2, 2]) = (0, 0, i2, i5, i1)

F(i5, 1, [2, 5]) = (0, i2, i5, i1, i4, 0, 0, 0)

Необходимо сделать одно уточнение - не всегда будет существовать поток для всей окрестности.

Такая ситуация возможна, когда заданный информационный элемент находится близко к началу или концу информационного потока, описывающего текст, т.е. элементов i - r-, …, i - 2, i - 1 и i + 1, i + 2, …, i + r+ может просто не существовать, или существовать только часть их, в зависимости от того, насколько близок элемент i к началу или концу текста.

Также такая ситуация возможна в том случае, когда циклы, формируемые потоком, многократно проходящим через i, имеют меньше информационных элементов, чем выбранное r.

В таких случаях будем принимать равным 0, все элементы, которые не определены в данной окрестности.

Обозначим множество всех информационных потоков относительно информационного элемента i для всех его вхождений в поток F:



Пример:

F = (i2, i5, i1, i4, i5, i2, i3, i8, i5, i1),

d(i5) = 3,

D(F, i5, [2, 2]) = ((0, i2, i5, i1, i4), (i1, i4, i5, i2, i3), (i3, i8, i5, i1, 0)).

На базе представленной модели выполним разработку метода и алгоритмов тематического анализа текста для решения двух основных задач данной работы:

1) тематической классификации текстовой информации;

2) вычисления степени тематической принадлежности текста к заданному классу.

КОРОТКО О РАЗРАБОТАННОЙ ПРОГРАММЕ