АЛГОРИТМЫ ПОИСКА В АССОЦИАТИВНО-ВЕРБАЛЬНЫХ СЕТЯХ ПСИХОЛИНГВИСТИЧЕСКИХ ЭКСПЕРИМЕНТОВ

А.В.Сиренко
(Московский государственный университет печати, г. Москва)

В статье рассматривается ассоциативно-вербальная сеть, сформированная в результате психолингвистического ассоциативного эксперимента. Указываются ключевые особенности эксперимента, оказывающие влияние на ассоциативно-вербальную сеть. Анализируется структура сети и ее количественные характеристики. В данном изложении, ассоциативно-вербальная сеть представлена в качестве направленного разреженного графа большой размерности. Приведены основные задачи, стоящие перед исследователем ассоциативно-вербальной сети, и методы их решения с использованием реляционных систем управления базами данных для хранения сети и выполнения базовых вычислительных операций. В конце статьи изложен подход к дальнейшему развитию хранения и поиска в ассоциативно-вербальной сети, рассмотрена задача анализа ее структуры на примере кластеризации.

1. АССОЦИАТИВНЫЙ ЭКСПЕРИМЕНТ

1.1. ФОРМАТ ЭКСПЕРИМЕНТА

Ассоциативно-вербальная сеть (далее АВС), о которой пойдет речь в данной статье, является результатом ассоциативного эксперимента, проведенного ранее [Черкасова, 2004]. В данном эксперименте респондентам предъявлялось некое слово – стимул, и предлагалось ответить на него первым же пришедшим на ум понятием, названным реакцией. Список стимулов был сформирован заранее и дополнялся в процессе эксперимента, реакция респондента могла быть произвольной и фиксировалась в текстовой форме. Также фиксировался гендерный признак респондента.

Ассоциативно-вербальная сеть представлена кортежами следующего вида:

{St, Rk, Count}, где

St – стимул в текстовой форме,

Rk – реакция в текстовой форме,

Count - число повторов ассоциации в процессе эксперимента.

Таким образом, моделью исходной АВС может быть направленный граф.

АВС подверглась некоторому преобразованию: из исходного множества кортежей были выделены таблицы узлов и связей согласно третьей нормальной форме организации баз данных. Для каждой связи была получена величина, отражающая вероятность перехода от некоторого стимула к реакции по данной связи. Обозначим ее как Prob (Probability – анг. вероятность), принимаемое значение от 0 до 1.

1.2. ОСОБЕННОСТИ АССОЦИАТИВНО-ВЕРБАЛЬНОЙ СЕТИ

Основной целью, преследуемой при формировании ассоциативно-вербальной сети, является фиксация представления носителя языка об окружающей действительности, имеющихся в ней отношениях и субъективном отношении респондента. Поэтому АВС должны иметь большую размерность, чтобы составлять базовое представление о языковой картине мира носителя языка

Всего узлов	103 349
Стимулов	6 624
Всего связей	457 534
Связей между стимульными узлами	228 038
Единичных связей(count = 1)	339 084

Таблица 1 Состав ассоциативно-вербальной сети

Число раз фиксации ассоциации в процессе эксперимента, характеризует ее устойчивость для носителя языка. Заметим, что 49% связей объединяет 6,4% узлов сети. Несомненно, методика проведения психолингвистического эксперимента, влияет на количественные показатели. В то же время, лишь 23% реакций были зафиксированы более чем единожды (предполагаются реакции, не являющиеся стимулами ни в каких связях). Можем сделать вывод, что сеть содержит некое ядро, окруженное множеством реакций, играющих роль выходов из сети.

2. АЛГОРИТМЫ АНАЛИЗА АВС

2.1. ПОСТАНОВКА ЗАДАЧИ

АВС может представлять интерес как с точки зрения анализа ее структуры, так и в качестве носителя информации о взаимосвязях между отдельными узлами сети. Характерным методом интеллектуального анализа сети как целого является кластеризация: определение неоднородностей сети, группировка узлов согласно их взаимной удаленности[1]. Таким образом, основной элементарной операцией, производимой с АВС, является вычисление путей между узлами сети. На пути может накладываться ограничение касательно их эффективности.

Поиск путей на графах является хорошо изученной областью вычислительных задач, так как эта модель отражает многие процессы и отношения действительности. Транспортные задачи, бизнес-процессы

[1] Здесь и далее под удаленностью понимается величина, обратная силе связи между узлами. Зависит от выбранной метрики сети.

предприятий и многое другое представимо в виде графов. В то же время, вычислительная сложность решения задач на графах (вычисления путей с заданными характеристиками, определение связных компонент и т.д.) зависит от вида графа: направленный или не направленный, ациклический или содержащий циклы.

2.2. БАЗОВЫЕ АЛГОРИТМЫ ПОИСКА С ИСПОЛЬЗОВАНИЕМ СУБД

Наиболее простым в организации с помощью реляционных баз данных является волновой алгоритм поиска. Предположим, нам необходимо найти минимальный путь от узла St_1 к узлу Rk_1(здесь связанные узлы считаем равноудаленными):

1. Выполняем поиск всех узлов, непосредственно достижимых из узла St_1.
2. Формируем из них множество M_1;
3. Если конечный узел присутствует в M_1 – задача выполнена. Иначе выполняем поиск всех узлов, достижимых из M_1 и формируем из них множество M_2.

Ниже представлено число узлов во множествах M_1-M_6 при поиске путей от стимулов «оружие», «лук» к реакции «арбалет» путей длиной до 6.

Номер шага	Число достижимых узлов
1	2
2	105
3	3607
4	42252
5	92526
6	99850

Таблица 2 Волновой поиск

Алгоритм потребует обхода связей в обратном направлении по множествам M_i от Rk_1 к St_1, для сбора информации о топологии путей. На практике такой подход оказывается более эффективным, чем классические поиск в глубину и ширину, так как большинство узлов во множествах Mi не входят в искомый путь, а хранение в последующем отбрасываемых путей вызывает значительный расход ресурсов.

Ассоциативно-вербальная сеть содержит циклы. Эмпирически определено, что в случае, если два узла сети достижимы, они достижимы за число шагов, не превышающее число 7. Несомненно, при заполнении АВС могут возникнуть слабосвязные[2] с остальным графом ассоциативные цепочки, нарушающие данное правило, но важно его выполнение в подавляющем большинстве запросов, так как с точки зрения практических задач больший интерес представляют короткие пути.

Была предложена модификация волнового алгоритма поиска, состоящая в поиске как от стимула к реакции, так и в обратном направлении, одновременно. Подобный двунаправленный алгоритм может балансироваться в процессе выполнения согласно соотношению узлов во множествах M_i и N_j, где M_i – множество достижимых узлов последнего шага в прямом направлении, N_j-множество достижимых узлов последнего шага в обратном направлении. В текущей итерации выполняется поиск в прямом направлении, при $M_i < N_j$, иначе - в обратном.

Приведем последовательность поиска при поиске путей от стимулов «оружие», «лук» к реакции «арбалет» путей длиной до 6, для сравнения с первоначальным алгоритмом (Таблица 2).

Направление	Номер шага	Число достижимых узлов
Прямое	1	2
Обратное	1	1
Обратное	2	2
Обратное	3	34
Прямое	2	105
Обратное	4	1419

Таблица 3 Двунаправленный волновой поиск

Модификация алгоритма позволила определить пути, совершив обход 0,6% числа связей, рассмотренных базовым волновых алгоритмом.

2.3. МОДИФИКАЦИЯ АЛГОРИТМОВ | ПОИСКА С УЧЕТОМ МЕТРИКИ СЕТИ

Приведенный выше алгоритм осуществляет поиск цепочек в сети безотносительно к близости узлов сети. В то же время важнейшим свойством ассоциации является число ее появления в процессе ассоциативного эксперимента, производной от которой мы определили ранее вероятность перехода от стимула к реакции согласно данной ассоциативной связи (Prob в кортеже параметров связи). Данный параметр служит мерой непосредственной достижимости узлов. Использование подобной метрики меняет подход к поиску путей, поскольку кратчайшим может быть путь, содержащий большее число промежуточных узлов.

Алгоритм поиска должен совершать обход графа подобно алгоритму поиска в глубину с тем отличием, что на каждом шаге поиск должен проводиться в направлении самого короткого пути. Очевидно, что предыдущий алгоритм можно считать частным случаем, в условиях равной удаленности связанных узлов.

Возможно группировать узлы в более общие структуры таким образом, чтобы понизить размерность сети. Назовем их «виртуальными узлами».

Подобная группировка будет эффективной, если:

1. Число связей внутри выделенной группы превосходит число связей между группой и внешней к ней частью сети, то есть происходит скрытие сложности графа.

2. Алгоритм поиска в определении порядка обхода сети руководствуется теми же правилами, которые используются при выделении виртуальных узлов. После захода в виртуальный узел, прежде всего, будет совершен обход узлов внутри виртуального узла.

[2] Согласно определению [Харрари 2006], стр.233.

3. Размеры выделенных подграфов достаточно велики, чтобы преимущества свертки превзошли затраты на усложнение структуры.

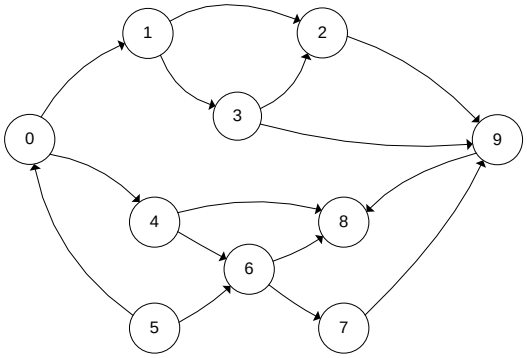

Рис. 1 Исходный граф сети

В общем случае, виртуальные узлы могут образовывать иерархии, увеличивая связность исходного графа. Задача выделения виртуальных узлов соответствует задаче четкой агрегативной кластеризации.

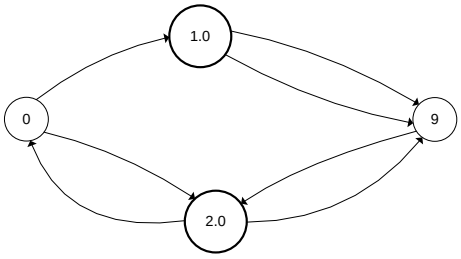

Рис. 2 Граф верхнего уровня после выделения виртуальных узлов

2.4. КЛАСТЕРИЗАЦИЯ СЕТИ

Основой любого метода кластеризации является определение расстояния между узлами. Расстояния, разумеется, умозрительного, в качестве меры похожести, общности объектов. В работе «Ассоциации информационных технологий» [Филиппович, Черкасова, Дельфт, 2002] в числе прочих задач, проводилась кластеризация компактной ассоциативной сети, ограниченной областью информационных технологий. Авторами была предложена мера близости стимулов через общность множеств их реакций.

В АВС ассоциативного эксперимента мы можем расширить метрику, учитывая также непосредственные связи между сравниваемыми узлами, одновременно сделав эту меру общей для всех узлов сети, не только стимульных.

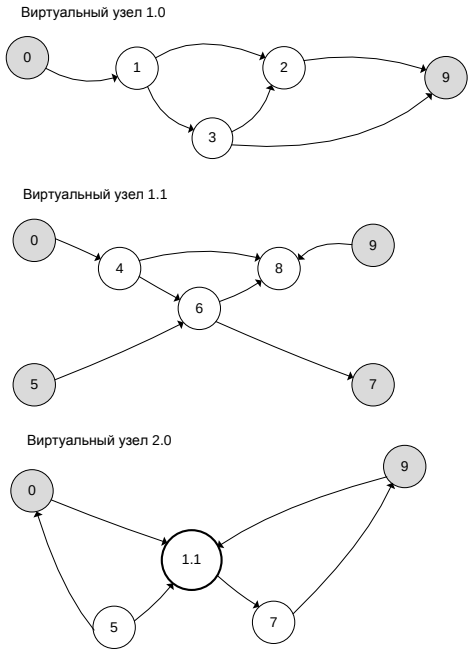

Рис. 3 Виртуальные узлы

К метрике расстояний предъявляются следующие требования:

1. Расстояние между элементами должно зависеть от списка общих реакций элементов и соответствующих вероятностей ассоциативной сети.

2. Расстояние между элементами должно зависеть от связей между ними в ассоциативной сети.

3. При отсутствии общих реакций и связей, расстояние должно равняться некоторому L_{max}, при увеличении связности элементов стремиться к L_{min}. Промежуточные значения должны располагаться от L_{min} до L_{max}.

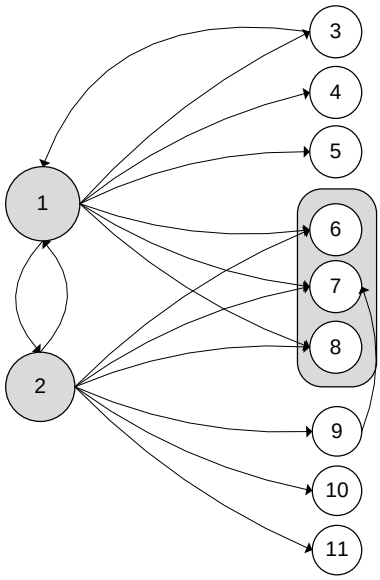

Рис. 4 Определение расстояния для кластеризации

Рис. 4 иллюстрирует определение расстояния между узлами 1 и 2. В общем случае узлы имеют совпадающие реакции (узлы 6,7,8), и непосредственные связи между собой.

Обозначим P_{ij} величину, соответствующую вероятности перехода в ассоциативной сети от узла i к узлу j по соответствующей связи (параметр Prob кортежа свойств ассоциации).

Пусть L_{ij} – искомое расстояние между узлами i и j. Тогда связность узлов i и j

$$F_{ij} = P_{ij} + P_{ji} + \sum_{k=1}^{N} \mathrm{Min}(P_{ik}, P_{jk}) \qquad (1)$$

где N соответствует числу узлов сети.

$$L_{ij} = L_{max} - \frac{F_{ij}}{2}(L_{max} - L_{min}) \qquad (2)$$

Коэффициент 2 в числителе формулы (2) установлен из соображений, что F_{ij} принимает значения от 0 до 2;

Сумма в формуле (1) на практике не производит просмотр всех узлов сети, поскольку для большинства узлов k связей P_{ik}, P_{jk} не существует. Для слабосвязных графов большой размерности эффективнее хранить связи в виде списков смежных узлов [Сэджвик, 2002].

Применительно к Рис. 4:

$F_{12} = P_{12} + P_{21} + \mathrm{Min}(P_{16},P_{26}) + \mathrm{Min}(P_{16},P_{26}) + \mathrm{Min}(P_{16},P_{26})$;

3. ЗАКЛЮЧЕНИЕ

Ассоциативно-вербальные сети являются мощным средством изучения мышления человека. В данной сфере существует множество как технических сложностей: требуемая большая размерность сети, наличие в ней шумов и неочевидных связей, трудоемкость сбора данных, так и сложностей, связанных с отсутствием формализованных методов анализа АВС. Это является следствием наших весьма общих представлений о человеческом мышлении. И, хотя АВС может быть представлена в виде графа с соответствующим математическим аппаратом, метрика сети, принципы определения расстояний между узлами, методы кластеризации и поиска в сети являются предметом эмпирических исследований.

В статье представлена модель ассоциативно-вербальной сети, ряд свойств которой уже определен, иные подлежат дальнейшему исследованию. Стоит заметить, что по мере накопления данных в АВС, ее свойства меняются, появляется возможность выделения шумов, поиска зависимостей.

Использование ассоциативных сетей сегодня широко распространено: при анализе продаж, естественно-языковых данных и во многих других приложениях. Рассмотренная ассоциативно-вербальная сеть, являясь результатом ассоциативного эксперимента, содержит специфический тип ассоциаций и потому, отличается от иных ассоциативных сетей методами анализа. Однако некоторые алгоритмические подходы, предложенные в статье, могут быть применены в графах аналогичного класса (размерности, направленности связей, весовых коэффициентов и т.д.).

СПИСОК ЛИТЕРАТУРЫ

1. *Г.А.Черкасова.* Формальная модель ассоциативного исследования.// Проблемы прикладной лингвистики. Выпуск 2. Сборник статей./ Отв. ред. Н.В.Васильева. –М.: «Азбуковник», 2004. – 400с.
2. *Ю.Н.Филиппович, Г.А.Черкасова, Д.Дельфт.* Ассоциации информационных технологий: эксперимент на русском и французском языках. С предисловием Н.В.Уфимцевой. М.: Изд-во МГУП, 2002. — 304 с.
3. *Сэджвик Роберт.* Фундаментальные алгоритмы на C++. Алгоритмы на графах: Пер. с англ. / Роберт Седжвик. – СПб: ООО «ДиаСофтЮП», 2002. – 496 с.
4. *Харрари Фрэнк.* Теория графов / Пер. с англ. и предисл. В.П. Козырева. Под ред. Г.П. Гаврилова. Изд. 3-е, стереотипное. – М.: КомКнига, 2006. – 296 с.