

**МЕДВЕДНИКОВА М.А.**, магистрант кафедры ИЗИ ВлГУ

**КОСТИНА Н.В.**, магистрант кафедры ИЗИ ВлГУ

## **АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ ПСИХОСЕМАНТИКИ И КОГНИТИВИСТИКИ НА ПРЕДМЕТ ИХ ПРИМЕНИМОСТИ ДЛЯ АНАЛИЗА МИКРОБЛОГОВ УЗЛА СОЦИАЛЬНОЙ СЕТИ**

Для исследования психосемантического профиля пользователя социальной сети была поставлена задача - создать методику оценки психосемантического профиля пользователя социальной сети и последующей ее автоматизации. В данной работе приведен аналитический обзор методов психосемантики и когнитивистики, используемых для анализа текстов на ЕЯ.

2 табл., 15 источник

В настоящее время одной из актуальных задач является оценка психосемантического профиля узла социальной сети, так как позволяет выявить особенности пользователя исходя из публикуемых сообщений.

Психосемантика – это направление языкознания, которое изучает связь между психологией и словом, между восприятием и значением слова.

Целью исследования является повышение эффективности предоставления контекстной рекламы, агитационных и прочих материалов, которые будут коррелировать с интересами пользователя.

Задача исследования заключается в создании методики оценки психосемантического профиля пользователя социальной сети и последующей ее автоматизации.

Результаты таких исследований помогут координировать социальные работы (например, агитационные, продвигаемые бизнес-идеи и т.д.) в социальных сетях.

Объектом анализа являются микроблоги узла социальной сети.

Предмет исследования – методы автоматизированной оценки и построения психосемантического профиля пользователя социальной сети.

Промышленные системы автоматической обработки текста используют следующую схему обработки текстов. К первой стадии (анализ отдельных слов) относится морфологический анализ (определение морфологических характеристик каждого слова — часть речи, падеж, склонение, спряжение и т.д.) и морфемный анализ (приставка, корень, суффикс и окончание); ко второй стадии — синтаксический анализ; к третьей — различные задачи семантического анализа (поиск фрагментов, формализация, реферирование и т.д.).

Анализ текста в микроблогах предполагается проводить на основе общепринятой схемы с некоторыми особенностями:

- Сбор информации со стены пользователя (временной интервал между публикуемыми сообщениями; средняя длина сообщения; процент сообщений без текста и другие).

- Морфологический анализ текста.
- Синтаксический анализ текста.
- Семантический анализ сообщений.
- Когнитивный анализ сообщений.
- Интеллектуальный анализ данных.

В сферу семантического анализа входит:

- Построение семантической интерпретации слов и конструкций;
- Установление "содержательных" семантических отношений между элементами текста,

которые уже принципиально не ограничены размером одного слова (могут быть больше или меньше одного слова).

Проанализируем некоторые известные методы семантического анализа с точки зрения применения для сообщений в социальной сети.

*Метод компонентного анализа* заключается в делении текста на семантические до тех пор, пока все необходимые слова не окажутся охваченными. Смысл предложения понимается в компонентном анализе как сумма смыслов входящих в него слов. Метод компонентного анализа удобен тем, что позволяет разграничить значения семантически близких слов. Но если исходить из принципа, что цель должна оправдывать средства, то такая функция в будущей системе не является столь важной. К недостаткам метода можно отнести то, что для его реализации необходимы большие словари, представленные в удобном виде.

*Метод семантических падежей.* Филмор принял идею компонентного анализа и развил ее вплоть до выделения семантически неделимых смысловых единиц. Согласно общепринятым взглядам центральной единицей смысла является предикат, под которым, как правило, подразумевается действие, выражаемое глаголом. Предикат имеет ряд атрибутов. По Филмору эти атрибуты имеют определенную семантическую роль. Он выделяет следующие роли: агент, объект, адресат, лицо, для которого совершается действие, пациент, результат, контрагент, источник, инструмент. Собственно эти роли и называются семантическими падежами Филмора. Основным результатом его исследований является пересмотр обычной схемы словарной статьи в толковом словаре. Он считает словарь основным средством задания семантических ролевых структур и правил их перевода в поверхностные структуры, общие с понятиями таких структур в российских исследовательских моделях управления Ю.Д. Апресяна. Недостатком данного метода является также необходимость в большой базе словарей.

*Ассоциативный эксперимент* является наиболее разработанной техникой психолингвистического анализа семантики. Он применяется в целях экспериментального исследования субъективных семантических полей слов, формируемых и функционирующих в сознании человека, а также семантических связей слов внутри семантического поля.

Данный метод приемлем для оценки статистического параметра «эмоциональное шкалирование», который позволяет оценить эмоциональность текстов.

Таблица 2 – Таблица для эксперимента по выявлению параметра эмоционального шкалирования

Шкала	Контрольные термины
1. Вдохновение	Воодушевление, Интерес, Увлечение, Подъём, Энтузиазм
2. Радость	Смех, Удовольствие, Удовлетворённость, Веселье
3. Равнодушие	Безучастность, Хладнокровие, Вялость, Апатия, Пофигизм
4. Страх	Робость, Ужас, Дрожь, Паника
5. Недовольство	Досада, Раздражение, Отвращение
6. Злость	Гнев, Негодование, Ярость, Бешенство
7. Неясность	Нечёткость, Непонятность, Туманность, Неточность, лывчатость, Двусмысленность
8. Расстройство	Огорчение, Печаль, Угнетение, Депрессия, Обида

«Эмоциональное шкалирование» даст представление о настрое текстов, публикуемых пользователем и соответственно об эмоциональной окраске текстов, лучше воспринимаемых пользователем.

К минусам данного метода можно отнести оценку результатов эксперимента (связано в том числе с наличием нецензурной лексики в микроблогах, исключением однокоренных слов и т.д.).

*Семантический дифференциал.* С помощью такого рода шкал можно получить от испытуемых оценки разного рода понятий.

Фактор оценки: веселый / грустный; хороший / плохой; полный / пустой; светлый / темный.

Фактор силы: длинный/ короткий; большой /малый; сильный / слабый; сложный / простой.

Фактор активности: новый /старый; теплый / холодный; быстрый / медленный; активный / пассивный.

Из плюсов можно выделить то, что эта методика получила широкое применение в теории массовой коммуникации и в рекламе (для выбора наиболее подходящих слов из ряда синонимов). Кроме того, семантический дифференциал применяется в исследованиях, связанных с восприятием и поведением человека, с анализом социальных установок и личностных смыслов.

*Построение семантических пространств.* Результатом является установление семантических связей значений, которые предстают в свернутом виде, как семантические

компоненты значений. Формой фиксации семантических связей являются семантические поля, тезаурусы и т. п. Данный метод будет хорошим дополнением к системе, а именно своего рода базой знаний в предметной области. Трудность заключается в получение тезауруса такого семантического пространства.

*Метод тематической классификации на основе модели структурного представления текста.* Суть метода заключается в том, чтобы выделить первичный набор ключевых элементов за счет количества их повторений в тексте и затем дополнить его контекстом, на основе анализа структурных особенностей мультиграфа, что вполне удовлетворяет цели исследования.

*Латентно-семантический анализ* - это теория и метод для извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных. Преимуществами метода ЛСА, по сравнению с банальным подсчетом встречаемости терминов в документах, является то, что решаются проблемы полисемии и синонимии. Программа распознает случаи использования одного и того же термина в различных значениях в том или ином контексте, а также определяет семантическую близость слов, описывающих одно и то же явление.

*Онтология* – набор понятий, сущностей определенной области знаний, ориентированный на многократное использование для различных задач. Из плюсов можно выделить то, что создание онтологии менее трудоемкая задача по сравнению с семантическими полями. Из минусов – необходимость больших словарей, представленных в удобном виде.

*Контент-анализ* (англ. content analysis; от content — содержание) — формализованный метод изучения текстовой и графической информации, заключающийся в переводе изучаемой информации в количественные показатели и ее статистической обработке. Характеризуется большой строгостью, систематичностью. Контент-анализ является частью семантического анализа. Особенность контент-анализа состоит в том, что он изучает документы в их социальном контексте. Данный метод особенно популярен, особенно при анализе политических изданий.

*Таблица 1 – Описание методов семантического анализа текстов на ЕЯ*

Методы семантического анализа	Особенности	Минусы	Комментарии
<i>Метод компонентного анализа</i>	Разграничивает значения семантически близких слов	Необходимость в больших словарях, представленных в удобном виде	Метод в будущей системе не является столь важным
<i>Метод семантических надежд</i>	Предикаты, используемые в методе, имеют атрибуты,	Необходимость в больших словарях представленных в	Метод трудоемкий для реализации

	которые определяют их семантическую роль	удобном виде	
<i>Ассоциативный эксперимент</i>	Применяется в целях экспериментального исследования субъективных семантических полей слов, формируемых и функционирующих в сознании человека	Сложна оценка результатов эксперимента (связано в том числе с наличием нецензурной лексики в микроблогах, исключением однокоренных слов и т.д.).	Метод будет играть вспомогательную роль
<i>Семантический дифференциал</i>	Применяется в исследованиях, связанных с восприятием и поведением человека, с анализом социальных установок и личностных смыслов	Для получения адекватной статистики метод должен тестироваться на большой аудитории людей	Метод может использоваться только как вспомогательный
<i>Построение семантических пространств</i>	Смысловые единицы имеют между собой описанные связи	Трудность заключается в получение тезауруса такого семантического пространства	Метод довольно трудоемкий, может использоваться как основной
<i>Метод тематической классификации на основе модели структурного представления текста</i>	Применяется для сравнения текстов между собой, реферирования и др.	Возможны ошибки из-за частых повторений в текстах слов, не являющихся ключевыми	Может использоваться как основной
<i>Латентно-семантический анализ</i>	Метод для извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных	Необходима большая база знаний	Решаются проблемы полисемии и синонимии
<i>Онтология</i>	Похожа на упрощенное семантическое пространство	Необходима большая база знаний	Можно использовать как основной метод
<i>Контент-анализ</i>	Изучает документы в их социальном контексте	<i>Выделение единиц счета:</i> они должны быть большими, чтобы выразить значение; ограниченными по объему настолько, чтобы не выражать много значений и т.д.	Популярный метод. Часто используется в области данного исследования

В рамках поставленной задачи предлагается при семантическом анализе текста использовать следующие методы: ассоциативный анализ, метод тематической классификации на основе модели структурного представления текста, семантический дифференциал, онтология, контент-анализ, латентно-семантический анализ.

Что касается методов когнитивной лингвистики, следует сказать, что в основе их лежит понятие «концепта» – некой абстрактной единицы, играющей ключевое значение. Концептом может быть принято любое понятие в зависимости от контекста.

Относительно самих методов, существует три базовых подхода к описанию концепта понятия или целого текста.

а) *Описание лексикографического значения.* Заключается в анализе многочисленных контекстов употребления концепта путём логической редукции семантических параметров. При этом не учитывается эмоциональный и психологический аспект восприятия.

б) *Описание психолингвистического значения.* Основан на ассоциативном эксперименте. При этом для сбора статистики ответов принимается только первая ассоциативная реакция, что позволяет сохранить чистоту эксперимента и исключает возможность ассоциации на предыдущую ассоциативную реакцию. Исключает формальный аспект построения текста.

в) *Комбинированный метод описания.* Учитываются и формальные параметры, и эффект восприятия информации. При моделировании концепта используют когнитивные признаки, а не отдельные значения. Затем описывается структура концепта на основании совокупности образа, информационного содержания. Результат формируется в своеобразный граф, от которого далее можно перейти к формализованной модели и формулам.

Авторы полагают, что для исследования следует применить третий тип описания концепта, так как в процессе пропаганды затрагиваются как формальные критерии непосредственно содержания сообщения, так и психологический аспект привлечения и удержания внимания пользователя.

Несколько слов о наборе методов *интеллектуального анализа данных data mining*. Главной особенностью данного подхода является необходимость обработки больших объёмов информации для получения корректного результата. Включает в себя множество различных методов анализа данных. Среди самых распространённых из них был проведён аналитический обзор.

а) *Дерева решений* – древовидная структура, представляющая совокупность решений. Деревья решений разбивают данные на группы на основе значений переменных, в результате чего возникает иерархия операторов «если – то», которые классифицируют данные.

б) *Ассоциативные правила* – всегда определяются на бинарных атрибутах. Проблема – алгоритм, который находит множество правил, будет вероятно находить также множество бесполезных правил. Когда как алгоритм, который находит только ограниченное число ассоциаций, без хорошей настройки, будет вероятно также пропускать множество значимой информации. Неоправданно сложен для исследования, высокие риски потерь полезных данных или появления лишних.

в) *Нейронные сети* – обычно состоит из множества узлов: входные узлы получают входные сигналы, выходные узлы задают выходные сигналы, а потенциально неограниченное число промежуточных слоёв содержат промежуточные узлы. Существует много различных архитектур для нейронных сетей, и каждая из них использует различные соединения и стратегии обучения, чтобы выполнить задачи. Проходит стадию обучения выполнять задачи и стадию выполнения задачи. Данный метод достаточно трудоёмкий для реализации и обучения системы в исследовании.

На данный момент среди методов анализа data mining подходящего не найдено. В дальнейшем предстоит продолжать изучать данную область, анализировать возможности уже изученных методов с целью подбора метода для исследования параметров микроблогов в соц.сетях.

*Основные проблемы понимания текста в обработке естественных языков таковы:*

1. Знание системой контекста и проблемной области и обучение этому системы. Например, из предложения «мужчина вошел в дом с красным портфелем» можно извлечь как представление о мужчине с красным портфелем, так и о доме с красным портфелем, если заранее не иметь в виду, что применительно к мужчинам употребление принадлежности портфеля гораздо вероятнее, чем применительно к дому.

2. Различная форма передачи синтаксиса (т.е. структуры) предложения в разных языках. Например, если синтаксическая роль слова (подлежащее, сказуемое, определение и т.д.) в английской речи во многом определяется положением слова в предложении относительно других слов, то в русском предложении существует свободный порядок слов и для выявления синтаксической роли слова служат его морфологические признаки (например, окончания слов), служебные слова и знаки препинания.

3. Проблема равнозначности. Предложения «длинноухий грызун бросился от меня наутек» и «заяц бросился от меня наутек» могут означать одно и то же, но могут иметь и разный смысл, например, если в первом случае имелся в виду длинноухий тушканчик.

4. Наличие в тексте новых для компьютера слов, например неологизмов. Самообучаемая система должна уметь «интуитивно» определить (возможно, и неправильно, но с возможностью в дальнейшем исправить себя) лексическую роль, морфологическую форму этого слова,

попробовать вписать его в существующую структуру знаний, наделить его какими-то атрибутами или выяснить все это в диалоге с оператором. Система, не способная к самообучению просто потеряет какое-то количество информации.

5. Проблема совместимости новой информации с уже накопленными знаниями. Новая информация может каким-то образом противоречить уже накопленной информации. Необходимо реализовать механизм, определяющий, в каких случаях нужно отвергнуть старую информацию, а в каких – новую.

6. Проблема временных противоречий. Так в предложении «я думал, что сверну горы» глагол в прошедшей форме «думал» сочетается с глаголом будущего времени «сверну».

7. Проблема эллипсов, то есть предложений с пропущенными фактически, но существующими неявно благодаря контексту словами. Например, в предложении «я передам пакет тебе, а ты – Ивану Петровичу» во второй части опущен глагол «передашь» и существительное «пакет»

## Литература

1. Чапайкина Н. Е. Семантический анализ текстов. Основные положения [Текст] / Н. Е. Чапайкина // Молодой ученый. — 2012. — №5. — С. 112-115.

2. Денисенко В.Н., Чеботарева Е.Ю. Современные психолингвистические методы анализа речевой коммуникации: Учеб. пособие. – М.: РУДН, 2008. – 258 с.

3. Четвериков А.А. (СПбГУ). Дифференцированное измерение эмоциональности текстов с помощью алгоритма PMI-IR (Статья из сборника «Материалы XXXVIII Международной филологической конференции 11–13 марта 2009 г.» Психолингвистика. Часть 1 / Отв. редактор Т. В. Черниговская. – СПб.: Факультет филологии и искусств СПбГУ, 2009. – 53 с.)

4. Чугреев В.Л., Санкт-Петербургский государственный электротехнический университет "ЛЭТИ" им. В.И. Ульянова (Ленина), Модель структурного представления текстовой информации и метод ее тематического анализа на основе частотно-контекстной классификации (диссертация), 2003 г.

5. Статья «Семантика. Методы семантического исследования» [Электронный ресурс]: [http://www.krugosvet.ru/enc/gumanitarnye\\_nauki/](http://www.krugosvet.ru/enc/gumanitarnye_nauki/)

6. Методика компонентного анализа [Электронный ресурс]: [http://www.lib.tpu.ru/fulltext/m/2011/IvanovaLS\\_VasilyevaSL](http://www.lib.tpu.ru/fulltext/m/2011/IvanovaLS_VasilyevaSL)

7. Леонид Святогор, Виктор Гладун «Семантический анализ текстов естественного языка: Цели и средства» [Электронный ресурс]: [http://www.foibg.com/ibs\\_isc/ibs-09/ibs-09-p06.pdf](http://www.foibg.com/ibs_isc/ibs-09/ibs-09-p06.pdf)



8. В. Ф. Петренко семантическое пространство субъективное [Электронный ресурс]: [http://epistemology\\_of\\_science.academic.ru](http://epistemology_of_science.academic.ru)

9. Захарова И.В., Стрюкова Г.А. «Семантический дифференциал как метод диагностики восприятия учащимися педагога» [Электронный ресурс]: <http://www.vzms.org/diagnostika.htm>

10. Яковлюк Александр Николаевич «Семантические аспекты взаимодействия простых многозначных и сложных однозначных слов современного немецкого языка» [Электронный ресурс]: [http://dibase.ru/article/29032010\\_yakovlukan/5](http://dibase.ru/article/29032010_yakovlukan/5). 2010 г.

11. Конспект лекций по дисциплине «Принципы и методы лингвистических исследований» [Электронный ресурс]: <http://rudocs.exdat.com/docs/index-57441.html>

12. Константин Селезнев «Обработка текстов на естественном языке» [Электронный ресурс]: <http://www.osp.ru/os/2003/12/183694/>

13. Статья «Методы формальной семантики» <http://yazykoznanie.ru/content/view/84/272/>

14. Статья «Методология когнитивного анализа семантики слова – это и новый взгляд на традиционную семантику, и шаг навстречу междисциплинарным исследованиям» [Электронный ресурс]: <http://rudocs.exdat.com/docs/index-350653.html>

15. Статья «Принципы семантико-когнитивного анализа концептов» [Электронный ресурс]: <http://zinki.ru/book/kognitivnaya-lingvistika/principy-semantiko/>