обладающие определенным значением признака, а в качестве метода расчета средней применим четвертый метод. Таким образом, дисперсия имеет следующее значение:

$$\sigma^2 = w(1-w) = 0{,}25$$

Для расчёта минимального объема выборки необходимо воспользоваться формулой:

$$n = \frac{t^2 w(1-w)N}{N\Delta^2 + t^2 w(1-w)}$$

В отличие от повторного отбора, в бесповторном необходимо знать приблизительную численность генеральной совокупности. К сожалению, органы статистики не публикуют информацию о количестве инновационный предприятий, поэтому данную величину необходимо оценить.

Таким образом, предложенный подход позволит подтвердить обоснованность гипотезы о необходимости реинжиниринга существующих локальных систем информационной поддержки и их интеграции с создаваемой системой распределенной информационной поддержки инноваций в Казахстане.

Литература

1. Инновационый менеджмент / И.Ю. Евграфова, Е.О. Красильникова. – М.: Окей- книга. - 2009. - 84 с.
2. Агарков С. А., Кузнецова Е. С., Грязнова М. О., Инновационный менеджмент и государственная инновационная политика, Академия Естествознания, 2011
3. Послание Президента Республики Казахстан «Стратегия «Казахстан-2050»: новый политический курс государства». 14 декабря 2012. Проверено 25.02.2018 из https://strategy2050.kz/ru/multilanguage/.
4. Анцибор А.В. Анкетирование, как метод экспресс-анализа инновационной деятельности, Евразийский союз ученых_28.11.15_11(20) – [Электронный ресурс].
5. www.instat.gov.al/media/2956/nace_rev11.pdf.

ФОРМАЛЬНАЯ МОДЕЛЬ ОЦЕНИВАНИЯ КАЧЕСТВА ЭКСТРАКЦИИ И ИДЕНТИФИКАЦИИ ЗНАНИЙ ИЗ СЛАБОСТРУКТУРИРОВАННОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

Хайрова Н.Ф., Мамырбаев О.Ж., Избасаров Е.Ж., Мухсина К.Ж.

Институт информационных и вычислительных технологий КН МОН РК
erlan_1081@mail.ru

Аннотация. *В данной статье рассматривается модель для определения каждого семантического поля по результатам выделения множества смысловых лингвистических единиц (полнотекстовых документов, сверхфразовых единств, абзацев и др.), соответствующих данной смысловой парадигме.*

Ключевые слова. *Модель оценивания качества, слабоструктурированная текстовая информация.*

Введение. Для оценки эффективности созданной информационно-лингвистической технологии экстракции и идентификации знаний из слабоструктурированной текстовой информации необходимо выделить метрики – совокупность объективно измеряемых показателей, характеризующих деятельность пользователей до и после внедрения данной технологии. К таким метрикам обычно относят время поиска пользователем информации по тому или иному вопросу и уровень знаний, извлеченных пользователями данной системы. При этом, в отличие от временных показателей, характеризующих длительность выполнения тех или иных процессов и достаточно просто поддающихся объективному измерению, метрика уровня знаний достаточно сложно поддается измерению. Тем не менее, основную ценность для социально-экономических организационных систем обычно представляют новые, скрытые, неявные и неформализованные знания, извлеченные из информационных потоков, позволяющие принимать новые нетрадиционные решения.

Постановка и алгоритм решения задачи. На сегодняшний день не существует стандартных бенчмарков для измерения качества извлечения из текстов знаний [7]. Обычно, для решения задачи измерения качества семантической классификации и информационного поиска используют *метод тестовых коллекций* [4,6], заключающийся в сравнении результатов работы исследуемой схемы на заранее определенных данных с оценками экспертов на тех же данных. В результате сравнения получается одно-двухкритериальная оценка эффективности. Поскольку понятия «эффективного излечения знаний», «качества знаний» не определены, количественная оценка результатов работы системы нетривиальна. Традиционный подход в подобных случаях – сравнение с «эталонным» результатом, – плохо применим из-за необходимости создания эталонного ответа для каждого конкретного набора электронных документов.

Для оценки работы системы используем алгоритм, для которого выводы, сделанные системой, согласуются с мнением экспертов. Для получения интегральных показателей качества работы системы идентификации знаний в слабоструктурированных текстовых информационных потоках применима методика усредненных метрик [2].

Обычно в поисковых системах точность вычисляется как отношение правильных выданных источников к общему числу выданных источников, а полнота – как отношение числа правильных выданных источников к общему числу правильных источников, существующих в системе [3,5]. Будем использовать показатели количественной оценки эффективности поиска и классификации, утвержденные межгосударственным стандартом по информации, библиотечному и издательскому делу [1]. Такими показателями являются: коэффициент точности – *precision*, коэффициент полноты – *recall* и коэффициент аккуратности *accuracy*.

Для определения перечисленных коэффициентов необходимо для каждого семантического поля по результатам выделения множества смысловых лингвистических единиц (полнотекстовых документов, сверхфразовых единств, абзацев и др.), соответствующих данной смысловой парадигме, определить: $n_{уу}$– число выделенных элементов, релевантных семантическому полю с точки зрения эксперта, $n_{уп}$– число выделенных элементов, не релевантных, с точки зрения эксперта, $n_{пу}$–

количество релевантных элементов, не выделенных системой, и n_{nn}– количество нерелевантных элементов, не выделенных системой (рис.1.).

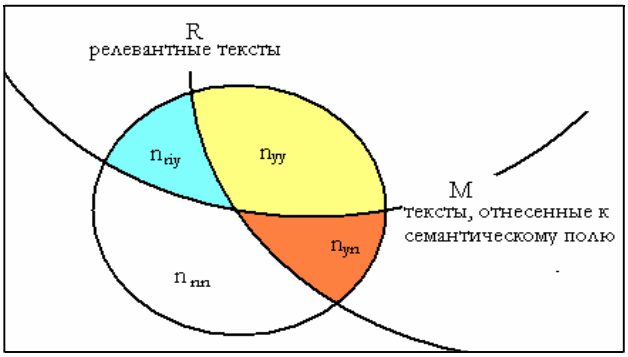

Рис. 1. Метрики оценки эффективности

При этом, если нет элементов, получивших с точки зрения эксперта определение *undefined*, сумма значений метрик равна количеству смысловых лингвистических элементов, поступивших на обработку:

$$D = n_{nn} + n_{ny} + n_{yn} + n_{yy},$$

где D – множество лингвистических элементов, поступивших в систему на обработку.

Отношение релевантности является субъективным, сложно определяемым и имеющим скорее психологическую природу понятием. Мы будем использовать определение релевантности [8], в котором релевантность зависит от четырех понятий *Relevance*(*IR, IN, C, T*), где *IR* – информационный ресурс, *IN* – информационная потребность, *C* – контекст и *T* – время. В нашей модели информационный ресурс представлен множеством лингвистических смысловых элементов уровня связного текста, поступившим на обработку $IR = D$. Информационную потребность можно разделить на неосознанную (истинную потребность) эксперта в знаниях, оперируя которыми эксперт решает некоторую информационную проблему, стоящую перед ним, и осознанную (внутреннее понимание реальной потребности). Переход между неосознанной и осознанной потребностью в знаниях вносит погрешность в вычисление эффективности работы подсистемы, основанной на интеллектуальных моделях.

Релевантность, определяемая неосознанной потребностью пользователя, лежит в основе определения коэффициента пертинентности информации [1]. Осознанная потребность интеллектуального поиска знаний определяет его полноту и точность. Осознанная потребность эксперта в знаниях, необходимых для решения некоторых задач, формируется в сфере мышления, но, сформировавшись в реальном контексте предметной области *C* и времени *T*, информационная потребность *IN* описывается средствами естественного языка.

При определении эффективности работы системы релевантность, т.е. соответствие связного текста крупной смысловой парадигме определяется экспертом по шкале “ *Relevance / irrelevant / undefined”* и показывает соответствие электронного текста некой локальной области знаний (или крупной смысловой парадигме).

Исходя из предположения, что система принимает решение о принадлежности к данной локальной области знаний каждого связного текста, вычисляем коэффициент

аккуратности как отношение правильно принятых решений системы к их общему числу:

$$accuracy = \frac{n_{yy} + n_{nn}}{n_{yy} + n_{ny} + n_{yn} + n_{nn}}$$

Ошибку вычисляем как отношение неправильно принятых системой решений к их общему числу:

$$error = \frac{n_{yy} + n_{nn}}{n_{yy} + n_{ny} + n_{yn} + n_{nn}}$$

В нашей модели полезно объединить точность и полноту в одной усредненной величине. Но при этом невозможно использовать среднее арифметическое, так как при стремящейся к нулю точности среднее арифметическое полноты и точности будет меньше 50 %. Будем опираться на среднее гармоническое точности и полноты, называемое мерой Ван Ризбергена, или *F-measure:*

$$F_1 - measure = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \times precision \times recall}{precision + recall}$$

При этом используемая в предложенной модели извлечения новых знаний из поступающей в систему слабоструктурированной текстовой информации метрика должна учитывать возможность использовать различные веса α для учета полноты и точности:

$$F_\beta - measure = \frac{1}{\alpha\frac{1}{precision} + \frac{(1-\alpha)}{recall}} = \frac{(\beta^2 + 1) \times precision \times recall}{\beta^2 \times precision + recall}$$

где $\alpha \in [0,1], \beta^2 = \frac{(1-\alpha)}{\alpha}, \beta \in [0, \infty]$.При значении коэффициентов $\alpha = 1/2$ или $\beta = 1F$-мера придает одинаковый вес полноте и точности и получается сбалансированная F_1-мера.

Если $0<\beta<1$- большее значение при расчете уделяется точности, а при $\beta > 1$– больший вес приобретает полнота. Так как в принятой системе расчета эффективности релевантность является субъективной, то эта субъективность будет перенесена на точность. В отличие от поисковой машины Интернет-поиска, в информационной системе, извлекающей знания для их последующего использования, точность приобретает первостепенное по отношению к полноте значение.

Заключение. Исходя из вышесказанного вданной модели можно объединить точность и полноту в одной усредненной величине. Но при этом невозможно использовать среднее арифметическое, так как при стремящейся к нулю точности среднее арифметическое полноты и точности будет не меньше 50 %. При этом используемая в предложенной модели извлечения новых знаний из поступающей в систему слабоструктурированной текстовой информации метрика должна учитывать возможность использовать различныевеса для учета полноты и точности.

Работа выполнена в рамках проекта ИРН AP 05131073 Методы и модели поиска и анализа криминально значимой информации в неструктурированных и слабоструктурированных текстовых массивах.

Литература

1. ГОСТ 7.73 – 96 SU. Поиск и распространение информации. Термины и определения. – Введ. 10.10.1996; принят Межгосуд. советом СНГ по стандартизации, метрологии и сертификации. – М.:Госстандарт.

2. Кураленок И. Оценка систем текстового поиска /И. Кураленок , И. Некрестьянов // Программирование. – 2002. – N 28(4). – С.226 –242.

3. Сегалович И. В. Методы сравнительного анализа современных поисковых систем и определения объема Рунета / И. В. Сегалович, Ю. Г. Зеленков, Д. О. Нагорнов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL'2006 : труды 8-й Всерос. науч. конф.; Суздаль, Россия, 2006. [Электронный ресурс] – Режим доступа: www.rcdl2006.unijar.ac.ru.

4. Шабанов В. И. Метод классификации текстовых документов, основанный на полнотекстовом поиске /В. И. Шабанов , А. М. Андреев // Труды РОМИП'2003. – СПб. : НИИ Химии СПб гос. ун-та, 2003. –С.52–71.

5. Bar-Yossef Z. and Gurevich. M. Efficient search engine measurements. In C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, editors. ACM. – Ban, Canada, 2007. – P.401–410.

6. Cormack G.V. A Efficient construction of large test collections / G. V.Cormack , C.R. Palmer ,C.L. Clarke // Proc. of the SIGIR'98– P.282–289.

7. Manning C., Schütze H. Foundations of Statistical Natural Language Processing. MIT Press,2000.

8. Mizzaro S. How many relevances in information retrieval? Proceeding of the Workshop «Information Retrieval and Human Computer Interaction», GIST Technical Report GR96-2 / Glasgow University. – Glasgow: The British Computer Society – P.57–60.

ПРИМЕНЕНИЕ ТЕХНОЛОГИИ ОТСЛЕЖИВАНИЯ ГЛАЗ В ИССЛЕДОВАТЕЛЬКИХ РАБОТАХ

Шокишалов Ж.М.

Институт информационных и вычислительных технологий КН МОН РК, Казахстан

e-mail*: jas_moderator@mail.ru*

Аннотация. *В последние несколько лет все большее внимание уделяется глазной биометрии. С одной стороны, это может быть связано с увеличением доступности глазных трекеров, т. е. устройств, способных измерять характеристики при движении глаз, такие как направление взгляда и размер зрачка. С другой стороны, функции и поведение глаз все больше и больше рассматриваются как потенциально более безопасные методы аутентификации, особенно при использовании в сочетании с традиционными методами проверки подлинности. В этой статье подробно*