РЕШЕНИЕ ПРОБЛЕМЫ ФОРМАЛЬНОЙ ОЦЕНКИ ЭФФЕКТИВНОСТИ ТЕХНОЛОГИЙ ИДЕНТИФИКАЦИИ ЗНАНИЙ В СЛАБОСТРУКТУРИРОВАННОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

Нина Хайрова, Наталья Шаронова, Дмитрий Узлов

***Аннотация**: В работе показана возможность использования интегральных количественных показателей полноты, точности и меры Ван-Ризбергена для оценки эффективности информационно-лингвистических технологий идентификации знаний в текстах. Обосновывается возможность использования метода тестовых коллекций для экспериментального подтверждения достоверности получаемых коэффициентов эффективности. В работе исследуется проблема максимизации надежности использования результатов, полученных по репрезентативной выборке, для выводов о генеральной совокупности текстовой коллекции. Рассмотрены процедуры использования выборочной доли признака как статистической характеристики для оценивания доли релевантных документов в генеральной совокупности. Предложен метод определения доверительного интервала для доли признака, основанный на подходе Вилсона, и метод определения необходимого объема релевантной выборки. Приведены примеры реализации предложенного подхода средствами Microsoft Excel.*

***Ключевые слова**: полнота, точность, релевантность, эффективность идентификации знаний, доверительный интервал, объем тестовой коллекции.*

ACM Classification Keywords: H.3.3 .Information Search and Retrieval, I.2.4. Knowledge Representation Formalisms and Methods, G.3. Probability and statistics – Statistical computing

Введение

Для оценки эффективности информационно-лингвистической технологии экстракции и идентификации знаний из слабоструктурированной текстовой информации необходимо определить 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 поддается измерению достаточно сложно. В то же время ясно, что основную ценность для социально-экономических организационных систем представляют, новые, скрытые, неявные и неформализованные знания, извлеченные, в том числе, из текстовых информационных потоков. Именно они позволяют принимать новые нетрадиционные решения.

На сегодняшний день не существует стандартных бенчмарков для измерения качества и эффективности технологии идентификации знаний, извлечённых из текстовых массивов. Обычно для определения эффективности технологий лингвистического процессора, используемых различными системами

семантической классификации и информационного поиска (Text Mining, Opinion Mining, Web Mining), используют метод тестовых коллекций [Шабанов, 2003; Cormack, 1998]. Суть данного метода заключается в сравнении результатов работы исследуемой системы на заранее определенных данных с оценками экспертов на тех же данных. В результате сравнения получается одно-двух-критериальная оценка эффективности. Но, поскольку для задач, связанных с экстракцией и идентификацией знаний, понятия „эффективного излечения знаний", „качества знаний" не имеют общепринятого определения, количественная оценка результатов работы системы не тривиальна. Традиционный подход в подобных случаях — сравнение с „эталонным" результатом — плохо применим из-за необходимости создания эталонного ответа для каждого конкретного набора электронных документов. Поэтому для оценки работы системы используется алгоритм, для которого выводы, сделанные системой, согласуются с мнением экспертов. Две основные, возникающие при этом проблемы заключаются в субъективности эксперта и в необходимом размере текстовой коллекции, позволяющем получить достоверный результат.

Под достоверностью понимается доказанная правильность того, что полученные в результате проведения эксперимента значения выполняются в определенных условиях для определенного класса объектов. Достоверность должна быть подтверждена верификацией, то есть повторением результатов в одних и тех же условиях при большом количестве проверок на разных объектах.

Общая постановка задачи

При использовании метода тестовых коллекций для оценки эффективности технологий идентификации знаний в слабоструктурированной текстовой информации возникают проблемы выделения интегральных количественных показателей оценки, учета реальных условий и количества исследований, а также проблема обоснования вывода о свойствах всех текстов коллекции (всей совокупности) по результатам выборочного метода экспериментального исследования. Иными словами при проведении проверки результатов информационно-лингвистических исследований на контрольных примерах возникает вопрос учета реальных условий и количества исследований.

Таким образом, прежде всего, необходимо обоснование использования в качестве усредненных метрик эффективности идентификации знаний в слабоструктурированной текстовой информации принятых показателей количественной оценки качества обработки текстовой информации, а именно – полноты, точности, шума и аккуратности, а также меры Ван Ризбергена.

Кроме того, в связи с тем, что при проведении экспериментальной проверки достоверности информационно-лингвистических технологий нецелесообразно или, скорее, практически невозможно в силу объективных причин, исследовать все тексты совокупности, то необходимо исследовать некоторую выборочную репрезентативную их часть. При этом возникает новый ряд проблем, связанный с необходимостью максимальной надежности использования результатов, полученных по выборке, для выводов о генеральной совокупности. Предлагается рассмотреть применение подхода, основанного на методах математической статистики (в частности, математической теории выборки), для определения объема экспериментальной выборки, необходимой для подтверждения достоверности разработанных информационно-лингвистических технологий, а также для нахождения погрешности оценивания выбранных показателей качества идентификации знаний в слабоструктурированной текстовой информации.

Интегральные показатели эффективности работы системы идентификации знаний

Для получения интегральных показателей эффективности работы системы идентификации знаний в слабоструктурированных текстовых информационных потоках применяем методики усредненных метрик.

Будем использовать показатели количественной оценки эффективности поиска и классификации, утвержденные межгосударственным стандартом по информации, библиотечному и издательскому делу [ISO 12620:2009]. Такими показателями являются: коэффициент точности — precision, коэффициент полноты — recall и коэффициент аккуратности accuracy, базирующиеся на субъективно определяемом понятии релевантности. Понятие релевантности является сложно определяемым и имеет, скорее, психологическую природу. Мы используем определение релевантности [Mizzaro, 1997], в котором релевантность зависит от четырех понятий Relevance (IR, IN, C, T), где IR — информационный ресурс, IN — информационные потребности, C — контекст и T — время. Информационный ресурс представлен множеством текстов коллекции, поступающим на обработку IR = D. Наибольшая субъективность при этом заключается в понятии информационной потребности, которую 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 двумя составляющими потребности вносит дополнительную погрешность в вычисление эффективности работы систем, основанных на знаниях, но только осознанная потребность эксперта в знании определяет полноту и точность работы системы. Дело в том, что именно осознанная потребность эксперта в знаниях, необходимых для решения некоторых задач, формируется в сфере мышления, и, сформировавшись в реальном контексте предметной области C и времени T, информационная потребность IN затем уже 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 или несоответствие электронного текста некой локальной области знаний (или крупной смысловой парадигме). Для определения коэффициентов полноты, точности и аккуратности необходимо для каждой области знаний эксперта определить: n_{yy} — число идентифицированных элементов (связных текстов или фрагментов связных текстов), релевантных области знаний эксперта, с его точки зрения, n_{yn} — число идентифицированных элементов, не релевантных области знаний, с точки зрения эксперта, n_{ny} — количество релевантных элементов, не идентифицированных системой, и n_{nn} — количество нерелевантных элементов, не идентифицированных системой (рис.1).

При этом, если нет элементов, получивших с точки зрения эксперта определение undefined, сумма значений метрик равна количеству элементов, поступивших на обработку: $D=n_{nn}+n_{ny}+n_{yn}+n_{yy}$, где D — множество элементов, поступивших в систему на обработку

Коэффициент точности определяется как:

$$precision = \frac{n_{yy}}{n_{yy} + n_{yn}}, \quad (1)$$

коэффициент полноты определяем как:

$$recall = \frac{n_{yy}}{n_{yy} + n_{ny}}. \quad (2)$$

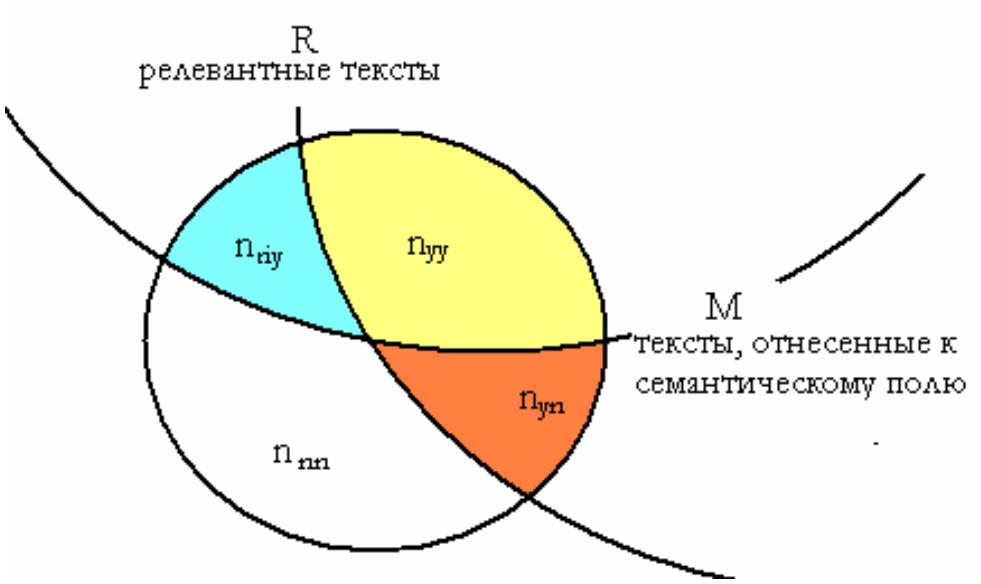

Рис. 1. Метрики оценки эффективности работы системы идентификации знаний

Для одновременного учета полноты и точности в одной усреднённой величине, с учетом различных весов α, можно использовать меру Ван Ризбергена или F-measure:

$$F_\beta - measure = \frac{1}{\alpha \frac{1}{precision} + (1-\alpha)\frac{1}{recall}} = \frac{(\beta^2+1)\times precision \times recall}{\beta^2 \times precision + recall} \quad (3)$$

где $\alpha \in [0,1]$, $\beta^2 = \frac{(1-\alpha)}{\alpha}$, $\beta \in [0,\infty]$. При значении коэффициентов α = 1/2 или β = 1 в F-мере полнота и точность имеют одинаковый вес и получаемая мера называется сбалансированной F_1-мерой:

$$F_1 - measure = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

Если $0<\beta<1$ — большее значение при расчете уделяется точности, а при $\beta>1$— больший вес приобретает полнота.

Выборочный метод исследования изучаемой совокупности

В связи с тем, что определяемые показатели эффективности работы системы идентификации знаний используют субъективно определяемое понятие релевантности той или иной смысловой парадигме или области знаний эксперта, достоверность значений показателей полноты, точности и меры Ван Ризбергена той или иной модели требует экспериментального подтверждения на совокупности текстов. Так как совокупность текстов, на которых реализуются модели, практически бесконечна, имеет смысл исследовать лишь часть объектов из изучаемой совокупности, т.е. осуществить так называемый выборочный метод исследования совокупности текстов и сделать обоснованные выводы о свойствах всей совокупности.

Одной из основных проблем выборочного исследования является получение репрезентативной выборки. Под репрезентативной выборкой понимается такая выборка, которая позволяет получить наиболее точную информацию о генеральной совокупности, а также не противоречит требованиям применимости вероятностных методов к обработке выборочных данных. Такая выборка образуется методом случайного отбора, то есть методом, при котором каждый элемент совокупности попадает в выборку случайным образом. Так как строгая реализация требований случайности отбора объектов в выборку требует разработки и применения специальных процедур и не всегда осуществима, на практике предполагается, что рассматриваемая выборка получена случайным отбором.

Так как в нашем случае отношение объема выборки к объему генеральной совокупности значительно меньше 5 – 10% (генеральной совокупностью любой модели обработки естественно-языковых текстов является стремящийся к бесконечному объему информационный текстовый поток), можно использовать математический аппарат теории возвратной выборки [Четыркин, 1982]. Кроме того, полученные результаты для возвратной выборки в ряде случаев можно перенести на соответствующий безвозвратный случай, вводя необходимый поправочный коэффициент.

Вторая проблема выборочного исследования заключается в проблеме оценки, связанной с тем, что выводы, делаемые на основе данных выборки, адекватно характеризуют только свойства выборки, а их перенос на свойства генеральной совокупности будет приводить к некоторой погрешности. Проблема оценки заключается в необходимости использования с максимальной возможной надежностью результатов, полученных по выборке для выводов о генеральной совокупности.

Для задачи оценки эффективности технологии идентификации знаний в слабоструктурированной текстовой информации будем оценивать долю признака (долю релевантных документов) в генеральной совокупности по соответствующей выборочной характеристике. При этом допускается, что любой рассматриваемый метод идентификации, как правило, не безошибочен: он может отнести к числу релевантных реально не релевантные документы, а также счесть не релевантными в действительности релевантные объекты.

Пусть доля признака в генеральной совокупности, которая показывает отношение числа релевантных документов к общему числу документов в генеральной совокупности, равна R. Выборочная оценка R_S доли R равна $R_S = M/N$, где, N – объем исследуемой возвратной выборки, а M – количество выявленных в ней рассматриваемым методом идентификации релевантных документов. Можно показать, что эта оценка удовлетворяет всем требованиям, предъявляемым к статистическим оценкам (состоятельность, несмещенность, достаточность и эффективность) [Четыркин, 1982]. Так как объекты случайно отбираются в выборку, то выборочная доля R_S может принимать любые значения в интервале [0;1], причем $R_S=0$, когда ни один релевантный документ не попал в выборку и $R_S=1$, если все документы в выборке релевантны.

Так как выборочная оценка R_S является точечной оценкой доли признака, то для нахождения погрешности приближения R оценкой R_S необходимо обратиться к интервальной оценке последней, то есть установить ошибку выборки. Поскольку R_S, а, следовательно, и ошибка выборки являются случайными величинами с одним и тем же распределением вероятностей, введем порождаемую конкретной точечной оценкой R_S интервальную оценку, в пределах которой с некоторой доверительной вероятностью P будет лежать доля признака генеральной совокупности.

Таким образом, доверительная вероятность P будет показывать вероятность того, что интервальная оценка содержит в себе неизвестную долю признака генеральной совокупности. Дополняющей вероятность P до 1 вероятностью α будем измерять риск выхода доли признака генеральной совокупности R за пределы интервальной оценки.

Поскольку в общем случае распределение величины R несимметрично, то интервальная оценка, или доверительный интервал, случайной величины R имеет вид:

$$P(R_S - E_1 < R < R_S + E_2) = 1 - \alpha, \quad (5)$$

где $[R_S - E_1; R_S + E_2]$ – доверительный интервал, $R_S - E_1$; $R_S + E_2$ – доверительные границы, $P = 1 - \alpha$ – доверительная вероятность, α – уровень значимости, или существенности.

В случае симметричного распределения R доверительный интервал также симметричен относительно величины R и имеет вид: $P(|R - R_S| < E) = 1 - \alpha$, где величина E – предельная ошибка, характеризующая точность выборки.

Обычно при таком подходе возникает несколько типов задач:

1) Определение доверительной вероятности по заданному доверительному интервалу и объему выборки;
2) Определение доверительного интервала по заданной доверительной вероятности и объему выборки;
3) Определение необходимого объема выборки по заданной доверительной вероятности и предельной ошибке.

В нашей задаче оценки эффективности технологий идентификации знаний в слабоструктурированной текстовой информации более актуальными являются определение доверительного интервала и определение необходимого объема выборки.

Определение доверительного интервала по заданной доверительной вероятности и объему выборки

Доверительный интервал для доли признака надо определять, строго говоря, базируясь на биноминальном законе распределения [Clopper, 1934]. Начиная с выборок объемом не менее 20, биномиальное распределение симметризуется и хорошо аппроксимируется нормальным распределением с параметрами: среднее $\langle R_S \rangle = R$, дисперсия $D(R_S) = R(1 - R)/N$, стандартное отклонение $\sigma(R_S) = [D(R_S)]^{1/2}$. При этом доверительный интервал может быть рассчитан по формуле $P(|R - R_S| < E_\alpha) = 2\Phi(Z_\alpha) = 1 - \alpha$, где $\Phi(Z_\alpha)$ – функция Лапласа. Предельная ошибка выборки находится при этом из равенства $E_\alpha = Z_\alpha \sigma(R_S)$.

В качестве величины доверительной вероятности обычно выбирают значение 0,95 (тогда уровень значимости α = 0,05). При этом $Z_{0,05} = 1,96$. Величина Z_α просто связана со статистической функцией Excel НОРМСТОБР(вероятность): Z_α = НОРМСТОБР(1 – α). С помощью этой функции может быть найдена величина Z_α при любой доверительной вероятности.

Теперь отталкиваясь от соотношения

$$|R - R_S| < Z_\alpha [R(1 - R)/N]^{1/2} \quad (6)$$

можно получить выражения для левой и правой границ доверительного интервала R, решая соответствующее квадратное уравнение относительно R [Wilson, 1927]. Использование подобного соотношения для адекватной оценки доверительных интервалов доли признака на малых выборках доказано статистиками [Brown, 2001; Garcia-Perez, 2005]. Тогда для левой границы доверительного интервала R_L имеем:

$$R_L = \frac{R_S + \frac{Z^2}{2N} - Z\left[\frac{R_S(1-R_S)}{N} + \frac{Z^2}{4N^2}\right]^{1/2}}{1+\frac{Z^2}{N}} \quad (7)$$

Для правой границы R_R получаем соответственно:

$$R_R = \frac{R_S + \frac{Z^2}{2N} + Z\left[\frac{R_S(1-R_S)}{N} + \frac{Z^2}{4N^2}\right]^{1/2}}{1+\frac{Z^2}{N}} \quad (8)$$

Опираясь на рекомендации нахождения доверительных границ, имеющиеся в работе [Agresti, 1998], получим значения доверительных границ более простым образом. Именно, переписав (6) в виде

$$Z_\alpha = |R - R_S|/[R(1 - R)/N]^{1/2}, \quad (6A)$$

и поместив формулу для Z_α в некоторую ячейку электронной таблицы, а какое-то, скажем 0,1, значение R в другую ячейку, воспользоваться сервисом ПОДБОР ПАРАМЕТРА MS-Excel, потребовав, чтобы в ячейке для Z_α подбиралось значение -1,96 (что соответствует доверительной вероятности 0,95), меняя параметр ячейки, содержащей R. Таким образом, будет найдена левая граница R_L доверительного интервала для R, а требуя подбора значения +1,96 в ячейке для Z_α, найдем соответствующую правую границу R_R.

Определение необходимого объема выборки

Для того чтобы определить объем нужной нам выборки при заданной доверительной вероятности и предельной ошибке, заменим в (6A) |R – RS| на E и разрешим получившееся уравнение относительно N. Тогда

$$N = [Z^2 R_S(1 - R_S)]/E^2 \quad (6B)$$

В соотношение (6B) входит выборочная доля R_S для определяемого еще неизвестного объема выборки. Поскольку эта доля неизвестна, разумно определить ее так, чтобы объем выборки N был максимальным (то есть годился при всех допустимых R_S). Нетрудно видеть, что максимум N как функции R_S достигается при R_S = ½, то есть $N_{MAX} = Z^2/4E^2$. Надо подчеркнуть, что если при исследовании возникают или априори имеются (по аналогии, из опыта) некоторые предположения о величине доли признака, то надо использовать эту величину в формуле (6B). Необходимый объем выборки будет при этом меньше максимального, что целесообразно.

Достаточно часто при нахождении долей признаков используется величина предельной ошибки E = 0,05. Используя электронные таблицы MS-Excel, рассмотрим следующий иллюстративный пример.

Пусть у нас имеется возвратная выборка объемом N = 10 объектов и доля признака в ней R_S = 0,9. Используя сервис „Подбор параметра”, получим значения левой и правой доверительных границ для доли признака в генеральной совокупности R при доверительной вероятности, равной 0,95: 0,59 < R < 0,98. Размах полученного доверительного интервала представляется излишне широким. Найдем максимальный объем выборки N_{MAX} для той же доверительной вероятности 0,95 (напомним, что при этом Z = 1,96) и предельной ошибке E = 0,05. Округляя рассчитанный требуемый объем выборки до целого вверх, имеем: N_{MAX} = 385. Вновь, как и ранее, используя сервис „Подбор параметра”, получаем новый более узкий доверительный интервал: 0,87 < R < 0,93. Это достигнуто ценой значительного роста требуемого объема выборки.

Изложенная выше схема расчета требуемого объема для возвратной выборки может быть перенесена на случай безвозвратной. При этом, правда, генеральная совокупность должна быть конечной и нужно знать ее объем $N_{ГС}$. Для указанного переноса необходимо лишь в использованную ранее формулу для выборочного стандартного отклонения возвратной выборки ввести корректирующий множитель $[(N_{ГС} - N)/(N_{ГС} - 1)]^{1/2}$: $\sigma_{ББ}(R_S) = \sigma_{ВВ}(R_S)[(N_{ГС} - N)/(N_{ГС} - 1)]^{1/2}$. Здесь$\sigma_{ВВ}$ и $\sigma_{ББ}$ стандартные отклонения для возвратной и безвозвратной выборок соответственно.

Действуя, как и ранее, получим соотношение для требуемого при заданной погрешности и доверительной вероятности объема безвозвратной выборки:

$$N_{ББ} = [Z^2 R_S(1 - R_S)N_{ГС}]/[E^2(N_{ГС} - 1) + Z^2 R_S(1 - R_S)] \quad (6C)$$

Вновь наибольшее значение объема выборки получается при R_S = ½, и $N_{ББМАХ} = (Z^2 N_{ГС})/[4E^2(N_{ГС} - 1) + Z^2]$. Это выражение может быть переписано в виде $N_{ББМАХ} = (N_{ВВМАХ}N_{ГС})/[N_{ГС} - 1 + Z^2/4E^2]$. Отсюда следует, что $N_{ББМАХ} < N_{ВВМАХ}$, если $Z^2/4E^2 - 1$ положительно, что заведомо имеет место при используемых обычно Z = =1,96 и E = 0,05. Так при объеме генеральной совокупности в 1000 объектов мы получим для максимального объема безвозвратной выборки (при тех же требованиях к точности и надежности) $N_{ББМАХ}$ = 278.

Формулу для доверительного интервала в случае безвозвратной выборки точно так же можно получить, модифицируя вышеуказанным образом выражение для стандартной ошибки. Проще всего это сделать, отправляясь от (1A), с использованием сервиса „Поиск решения” MS-Excel. При этом получаем:

$$Z_\alpha = |R - R_S|/\{[R(1 - R)(N_{ГС} - N)]/[N(N_{ГС} - 1)]\}^{1/2} \quad (6D)$$

Для прежних исходных данных, $N_{ГС}$ = 1000, N = 10, и доверительной вероятности 0,95 с точностью до второго знака после запятой получим прежний доверительный интервал: 0,59 < R < 0,98. При большей точности доверительный интервал для безвозвратной выборки незначительно уже, чем для возвратной, поскольку объем выборки составляет всего 1% от объема генеральной совокупности. Для безвозвратной выборки максимального размера, обеспечивающей заданную точность и надежность, доверительный интервал с точностью до второго знака после запятой совпадает с аналогичным результатом для возвратной выборки: 0,87 < R < 0,93. Подчеркнем, что требуемый объем безвозвратной выборки при этом на 28% меньше.

Выводы

Таким образом, в работе обосновывается использование в качестве усредненных метрик эффективности идентификации знаний в слабоструктурированной текстовой информации принятых показателей

количественной оценки качества обработки текстовой информации, а именно – полноты, точности, шума и аккуратности, а также меры Ван Ризбергена. Для подтверждения достоверности разрабатываемых технологий применяется метод тестовых коллекций, при использовании которого необходимо решать проблему максимизации надежности использования результатов, полученных по тестовой коллекции, для выводов о генеральной совокупности всех исследуемых текстов. В работе рассмотрено применение подходов математической статистики для определения погрешности оценивания выбранных показателей качества идентификации знаний, а именно, использование методов определения доверительного интервала для доли признака и методов определения необходимого объема релевантной выборки в зависимости от заданной погрешности и доверительной вероятности.

Литература:

[Agresti, 1998] Agresti A., B. Coull A. Approximate is better than exact for interval estimation of binomial proportions // American statistician. – 1998. – N 52. – C. 119–126.

[Brown, 2001] L. D. Brown, T. T. Cai, A. Dasgupta. D. Interval estimation for a binomial proportion // Statistical science. – 2001. – N 2. – P. 101–133.

[Clopper, 1934] Clopper C. J., E. S. Pearson. The use of confidence or fiducially limits illustrated in the case of the binomial // Biometrika. – 1934. – N 26. – P. 404–413.

[Cormack, 1998] 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.

[Garcia-Perez, 2005] Garcia-Perez M. A. On the confidence interval for the binomial parameter // Quality and quantity. – 2005. – N 39. – P. 467–481.

[ISO 12620:2009] "ISO 12620:2009 - Terminology and other language and content resources -- Specification of data categories and management of a Data Category Registry for language resources".iso.org. 2011. Retrieved 9 November 2011.

[Mizzaro, 1997] Mizzaro S. Relevance: The whole history. Journal of American Society for Information Science. — 1997. — V.48. — Is. 9 — P. 810-832

[Wilson, 1927] Wilson E. B. Probable inference, the law of succession, and statistical inference // Journal of American Statistical Association. – 1927. – N 22. – P. 209–212.

[Медик, 2007] Медик В. А., Токмачев М. С. Математическая статистика в медицине. – М.: Финансы и статистика. 2007.

[Четыркин, 1982] Четыркин Е.М., Калихман И.Л. Вероятность и статистика. М.: Финансы и статистика. 1982

[Шабанов, 2003] 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.

Сведения об авторах

Нина Хайрова *– профессор кафедры интеллектуальных компьютерных систем Национального технического университета „Харьковский политехнический институт", ул. Фрунзе, 21, Харьков, 61002, Украина; e-mail: nina_khajrova@yahoo.com*

Научные интересы: искусственный интеллект, идентификация знаний из текстов, Text Mining, Opinion Mining, Web Mining, Natural language processing, искусственный интеллект

Наталья Шаронова *- профессор, заведующий кафедрой интеллектуальных компьютерных систем Национального технического университета „Харьковский политехнический институт", ул. Фрунзе, 21, Харьков, 61002, Украина; e-mail: nvsharonova@mail.ru*

Научные интересы: искусственный интеллект, математическое моделирование, автоматизированные библиотечные системы

Узлов Дмитрий *– соискатель кафедры интеллектуальных компьютерных систем Национального технического университета „Харьковский политехнический институт", ул. Фрунзе, 21, Харьков, 61002, Украина; e-mail: poputcik@mail.ru*

Научные интересы: системы автоматической обработке текстов на естественном языке, экстракция и идентификация знаний, искусственный интеллект

Solution of the Problem of Formal Evaluation of Effectiveness of the Technology Knowledge Identification in Semistructured Text Information

Nina Khairova, Nataliya Sharonova, Dmytro Uzlov

Abstract: The traditional approach (the comparison with a "reference" result) for evaluating quality of the technology to identify knowledge extracted from text arrays is badly applicable out of a need to create the reference answer for each specific set of electronic documents. In this paper we show that integral quantitative coefficients of recall, precision and F-measure can be used to assess effectiveness of linguistic technologies of knowledge identification in texts. Justifying the possibility of using the test collections method for the experimental validation of obtained efficiency coefficients, we propose the use of the approach based on mathematical statistics methods. The procedures of using sampling fraction of the indicator as a characteristic of evaluating the proportion of relevant documents in the general population are reviewed. The paper shows the argumentation to the fact that, in important practical cases of text collection samples, asymmetry of a confidence interval at the binomial distribution can be overcome by approximated transition to the normal distribution. We also propose the methods of determining the confidence interval for the indicator fraction that are based on Wilson approach, and the method of determining the required size of the relevant sample depending on the specified error and confidence probability as well.

Key worlds: evaluation of effectiveness, semistructured text information, test collections method, size sample