УДК 005.6

ИНСТРУМЕНТАРИЙ АВТОМАТИЗИРОВАННОЙ АГРЕГАЦИИ ПОТРЕБИТЕЛЬСКИХ ОЦЕНОК КАЧЕСТВА ПРОДУКЦИИ В СИСТЕМЕ МЕНЕДЖМЕНТА КАЧЕСТВА МАШИНОСТРОИТЕЛЬНОГО ПРЕДПРИЯТИЯ С ПОЗИЦИИ ОБЕСПЕЧЕНИЯ ПРИНЯТИЯ РЕШЕНИЙ НА ОСНОВЕ СВИДЕТЕЛЬСТВ

© 2026 В.Н. Пиунов

Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара, , Россия

Статья поступила в редакцию 14.09.2026

В статье описана архитектура программного сервиса агрегации обратной связи и кодификации воспринимаемого потребителями качества на основе гибридного поиска. Обоснован выбор лексического алгоритма BM25 и трансформерной модели BERT в режиме Bi-Encoder для параллельного поиска кандидатов. Предложена схема слияния ранкингов методом Reciprocal Rank Fusion (RRF) с последующим реранкингом Cross-Encoder. Приведены формулы BM25 и RRF, описана архитектура конвейера обработки, выполнен сравнительный анализ компонентов системы.
Ключевые слова: семантический поиск, гибридный поиск, BM25, BERT, Sentence-BERT, реранкинг, Reciprocal Rank Fusion, агрегация обратной связи, воспринимаемое качество, менеджмент качества.

DOI: 10.37313/1990-5378-2026-28-5-147-150
EDN: RXKFKZ

ВВЕДЕНИЕ

Современные предприятия сталкиваются с непрерывным ростом объёма неструктурированных обращений от заказчиков: жалоб, предложений, описаний неисправностей, запросов на улучшение функциональности. Обращения поступают из нескольких каналов связи, формулируются на естественном языке с широкой вариативностью терминологии, профессионального жаргона и степени детализации. Обработка этих данных вручную трудозатратна и её результаты с трудом поддаются воспроизводимому агрегированию в рамках систем менеджмента качества [1-4].

В работе Антоновой Н.А., посвящённой инструментам оценки воспринимаемого потребителями качества, была концептуально предложена идея создания лингвистического анализатора, который позволил бы правильно вскрывать сущность потребительских оценок качества продукции с использованием кодифицированных жалоб и предложений в качестве основы для автоматизированной обработки обращений. В рамках дальнейшего развития предложенной концепции в настоящем исследовании описан процесс агрегации обращений и классификации неструктурированных запросов, реализованный в виде специализированного сервиса агрегации. Данный программный компонент принимает зарегистрированные записи из централизованной базы данных, формирует семантически однородные группы и присваивает им коды тематического классификатора. Основу сервиса составляет алгоритм поиска ближайших соседей по смыслу, от точности которого зависит качество кластеризации и, в конечном счёте, кодификатора воспринимаемого потребителями качества [4].

В статье описываются технологические компоненты системы и её архитектура. В центре исследования - комбинация лексического и семантического поиска с обоснованием применением этой схемы для применения преимуществ этих двух подходов.

Постановка задачи и требования к архитектуре

Описываемый сервис должен обладать следующими функциями:

1. По вновь поступившей задаче находить в накопленной базе наиболее близкие по смыслу существующие записи.

2. На основе найденных соседей определять тематическую принадлежность запроса и присваивать ему код классификатора.

3. При появлении новой темы сигнализировать о необходимости расширения кодификатора.

Требования к архитектуре вытекают из природы входных данных:

Разнородность лексики. Разные заказчики по-разному описывают одну проблему: один - через симптом, другой - через последствие. Поиск должен обнаруживать семантически идентичные обращения вне зависимости от формулировки.

Наличие точных идентификаторов. Технические запросы содержат коды компонентов, артикулы, номера версий, аббревиатуры (ОКС, ГОСТ). Для таких запросов принципиально важно точное лексическое сопоставление [1].

Пиунов Владислав Николаевич, аспирант. E-mail: Taimkraft@mail.ru,

Масштабируемость. По мере накопления базы количество задач может исчисляться десятками тысяч; поиск должен выполняться за время, приемлемое для интерактивного режима работы.

Интерпретируемость. Результаты поиска должны поддаваться аудиту: специалист по качеству должен иметь возможность проверить, почему данная задача была отнесена к конкретной теме.

Ни лексический, ни семантический поиск в отдельности не удовлетворяют всем перечисленным требованиям одновременно. Это обусловливает применение гибридной схемы.

Лексический поиск: алгоритм BM25

Алгоритм BM25 (Best Match 25, также известен как Okapi BM25) разработан Робертсоном и Спэрк-Джонс в 1976 году и по сей день остаётся де-факто стандартом лексического ранжирования. В отличие от простого поиска по ключевым словам, BM25 учитывает:

1. частоту термина в документе с насыщением (убывающая отдача от повторений);
2. обратную частоту документа (редкие термины получают больший вес);
3. нормализацию по длине документа [5].

Оценка релевантности документа D запросу Q, содержащему термины q_1,...,q_n, вычисляется по формуле:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)}, \quad (1)$$

где f(q_i,D) - частота термина q_i в документе D; $|D|$ — длина документа в словах; avgdl - средняя длина документа в коллекции; $k_1 \in [1.2; 2.0]$ - параметр насыщения частоты термина; $b \in [0; 1]$ - параметр нормализации длины, обычно b=0.75. [5]

Обратная частота документа рассчитывается как:

$$\text{IDF}(q_i) = \ln\left(\frac{N-n(q_i)+0.5}{n(q_i)+0.5}+1\right), \quad (2)$$

где N – общее число документов в коллекции; $n(q_i)$ – число документов, содержащих термин q_i. Прибавление единицы под логарифмом предотвращает отрицательные значения IDF для высокочастотных терминов. [5]

BM25 обеспечивает высокую точность при поиске задач, содержащих точные технические идентификаторы, аббревиатуры и коды: это критически важно для технических обращений машиностроительных предприятий. При этом BM25 не способен улавливать семантическую близость разноформулированных запросов, что ограничивает его применение. Для работы с русскоязычными текстами необходима лемматизация (приведение слов к нормальной форме) до построения индекса – иначе флективные формы одного слова воспринимаются как разные термины.

Семантический поиск: BERT и Sentence-BERT

BERT (Bidirectional Encoder Representations from Transformers) - трансформерная языковая модель, предложенная командой Google AI в 2018 году. Ключевой особенностью является двунаправленный механизм самовнимания (Self-Attention): каждый токен при кодировании учитывает полный контекст слева и справа. Это принципиально отличает BERT от ранних рекуррентных моделей, обрабатывавших текст последовательно.

Предобучение BERT включает два этапа: Masked Language Modeling (MLM) - предсказание замаскированных токенов, и Next Sentence Prediction (NSP) - предсказание следования двух предложений. Такое предобучение на масштабных корпусах позволяет модели усваивать семантические зависимости, которые используются при последующей дообучении на конкретных прикладных задачах [6].

Оригинальный BERT при использовании в режиме CrossEncoder (оба текста подаются на вход одновременно) обеспечивает высокую точность, однако непригоден для поиска по большим базам: для ответа на каждый запрос требуется выполнить N прямых проходов через модель, что делает систему нерабочей при $N \gg 1000$.

Sentence-BERT (SBERT) решает эту проблему посредством архитектуры Bi-Encoder: два текста кодируются независимо в векторы фиксированной размерности, а их близость определяется косинусным расстоянием [7]:

$$\text{sim}(\mathbf{u},\mathbf{v}) = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\cdot\|\mathbf{v}\|}. \quad (3)$$

Благодаря этому все документы базы кодируются единожды и сохраняются в векторной базе данных. При поступлении нового запроса кодируется только он, а поиск ближайших соседей осуществляется за O(logN) с помощью индексов типа HNSW. Bi-Encoder используется как первичный ретривер – быстро отбирает k·m кандидатов. Cross-Encoder применяется на втором этапе как переранжировщик – перечитывает пару «запрос, документ» совместно и выдаёт точный показатель релевантности, но только для ограниченного числа кандидатов

Гибридный поиск: слияние методов алгоритмом Reciprocal Rank Fusion

Чистый семантический поиск уязвим для точных запросов по идентификаторам: вектор текста «договор №17-2024» будет иметь высокую косинусную близость к векторам других договоров вне

зависимости от их номера, поскольку кодировщик понимает «смысл», а не точную строку. Чистый лексический поиск, напротив, не находит семантически близкие документы с отличной формулировкой. Экспериментально показано, что гибридная схема BM25 + семантический поиск обгоняет каждый из компонентов в отдельности: по метрике Top-1 hit rate прирост от добавления BM25 к семантическому поиску составляет около 17 процентных пунктов.

Алгоритм Reciprocal Rank Fusion (RRF) предложен Cormack, Clarke и Buettcher в 2009 год. RRF объединяет несколько рейтингов, не требуя нормализации сырых оценок и не нуждаясь в обучении.

Итоговый рейтинг вычисляется по формуле:

$$\mathrm{RRF}(d) = \sum_{r \in R} \frac{1}{k + \mathrm{rank}_r(d)}, \quad (4)$$

где R - множество рейтингов; $\mathrm{rank}_r(d)$ – позиция документа d в рейтинге r; k – сглаживающий параметр.

Данный алгоритм часто применяется в промышленных поисковых системах. Например, в Elasticsearch, OpenSearch и Odrant. [8]

Архитектура сервиса и выбор технологий

Процесс обработки обращений включает четыре основных этапа:

Этап 0. Подготовка. Входящий текст задачи очищается от служебных символов и нормализуется. Для русскоязычного текста выполняется лемматизация. Нормализованный текст передаётся параллельно в BM25-индекс и в Bi-Encoder.

Этап 1. Поиск кандидатов. BM25 выполняет лексический поиск по инвертированному индексу и возвращает рейтинг топ-K документов. Bi-Encoder преобразует запрос в вектор и передаёт его в Qdrant, который использует HNSW-индекс для поиска ближайших соседей по косинусной близости [5-7].

Этап 2. RRF-слияние. Оба рейтинга объединяются по формуле (4).

Этап 3. Повторное ранжирование. Первые 50 документов объединённого рейтинга передаются на вход переранжировщика. Модель читает каждую пару (запрос, документ) и выдаёт точное значение релевантности. Итоговый рейтинг формируется по убыванию значения [8].

Этап 4. Тематическая агрегация. Топ-K документов анализируются с точки зрения их тематической принадлежности: если большинство ближайших соседей имеют один и тот же код кодификатора, новая запись получает тот же код. Если соседи относятся к различным темам или порог уверенности не достигнут, запись маркируется как потенциально новая тема для последующей верификации экспертом.

В таблице 1 приведено сопоставление технологических компонентов сервиса.

Таблица 1 – Сопоставление технологических компонентов сервиса

Компонент	Выбранная реализация	Альтернативы	Обоснование выбора
Лексический поиск	BM25 (rank_bm25)	TF-IDF, BM25+	BM25 превосходит TF-IDF по точности на разнородных корпусах; BM25+ даёт аналогичный результат, более сложен в реализации
Семантическое кодирование	DeepPavlov/rubert-base-cased, дообученный на MS MARCO RU	multilingual-e5, LaBSE	Специализированная русскоязычная модель; меньше вычислительных затрат по сравнению с мультиязычными
Векторная база данных	ChromaDb	Elasticsearch, FAISS, Weaviate	Поддержка гибридного поиска (sparse + dense), API для фильтрации по метаданным, интеграция с FastAPI
Слияние ранкингов	RRF (k = 60)	Linear combination, CombSUM	Не требует нормализации; не нуждается в обучении; устойчив к распределению скоров
Реранкер	Cross-Encoder	MonoT5, mxbai-rerank	Русскоязычная модель; достаточная точность для задач текстовой классификации
API-слой	FastAPI (Python)	Flask, Django	Асинхронная обработка, автогенерация OpenAPI-документации

ВЫВОД

Разработанная архитектура сервиса агрегации обратной связи обеспечивает использование и развитие существующих кодификаторов воспринимаемого потребителями качества посредством четырёхэтапного конвейера: лемматизация, BM25 поиск, RRF-слияние и переранжирование. Обоснован выбор технологий для каждого компонента.

Предложенная архитектура является основой программной реализации сервиса и может быть адаптирована для задач агрегации обратной связи на предприятиях различных отраслей.

СПИСОК ЛИТЕРАТУРЫ

1. *Гречников, Ф.В.* Научно-технический обзор современных подходов к применению цифровых технологий и искусственного интеллекта в управлении жизненным циклом продукции / Ф.В. Гречников, А.С. Клентак, В.Н. Пиунов, В.И. Ушаков // Известия Самарского научного центра Российской академии наук. – 2025. – Т. 27. – № 6.
2. *Антонова Н.А.* Совершенствование инструментов оценки и мониторинга воспринимаемого потребителями качества автомобилей в эксплуатации: дис. … канд. техн. наук: 2.5.22 / Антонова Наталья Алексеевна. – Самара, – 2024.
3. *Клентак А.С.* Разработка методологии управления конкурентоспособностью автопроизводителя в условиях промышленной сборки: дис. докт. техн. наук: 2.5.22 / Клентак Анна Сергеевна. – Самара, 2025.
4. *Клентак, А.С.* Концепция методологии прогнозирования востребованного потребителями качества продукции машиностроения при выходе компании-производителя на новые конкурентные рынки / А.С. Клентак, В.Н. Козловский, А.В. Гусев // Известия Самарского научного центра Российской академии наук. – 2024. – Т. 26. – № 1(117). – С. 5-14.
5. *Robertson S., Spärck Jones K.* Relevance weighting of search terms // Journal of the American Society for Information Sciences. 1976. № 27(3). С. 129–146.
6. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proc. NAACL. 2019.
7. *Reimers N., Gurevych I.* Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proc. EMNLP. 2019.
8. *Cormack G.V., Clarke C.L.A., Buettcher S.* Reciprocal Rank Fusion outperforms Condorcet and individual rank learning methods // Proc. 32nd ACM SIGIR. 2009. С. 758–759.

TOOLS FOR AUTOMATED AGGREGATION OF CONSUMER EVALUATIONS OF PRODUCT QUALITY IN THE QUALITY MANAGEMENT SYSTEM OF A MACHINE-BUILDING ENTERPRISE FROM THE PERSPECTIVE OF SUPPORTING DECISION-MAKING BASED ON TESTIMONIALS

© 2026 V.N. Piunov

Samara National Research University named after Academician S.P. Korolyov, Samara, Russia

This article describes the architecture of a software service for feedback aggregation and codification of consumer-perceived quality based on hybrid search. The choice of the BM25 lexical algorithm and the BERT transformer model in Bi-Encoder mode for parallel candidate search is justified. A scheme for merging rankings using Reciprocal Rank Fusion (RRF) followed by Cross-Encoder reranking is proposed. The BM25 and RRF formulas are presented, the processing pipeline architecture is described, and a comparative analysis of the system components is performed.

Keywords: semantic search, hybrid search, BM25, BERT, Sentence-BERT, reranking, Reciprocal Rank Fusion, feedback aggregation, perceived quality, quality management.

DOI: 10.37313/1990-5378-2026-28-5-147-150
EDN: RXKFKZ

REFERENCES

1. *Grechnikov, F.V.* Nauchno-tekhnicheskij obzor sovremennyh podhodov k primeneniyu cifrovyh tekhnologij i iskusstvennogo intellekta v upravlenii zhiznennym ciklom produkcii / F.V. Grechnikov, A.S. Klentak, V.N. Piunov, V.I. Ushakov // Izvestiya Samarskogo nauchnogo centra Rossijskoj akademii nauk. – 2025. – T. 27. – № 6.
2. *Antonova N.A.* Sovershenstvovanie instrumentov ocenki i monitoringa vosprinimaemogo potrebitelyami kachestva avtomobilej v ekspluatacii: dis. … kand. tekhn. nauk: 2.5.22 / Antonova Natal'ya Alekseevna. – Samara, – 2024.
3. *Klentak A.S.* Razrabotka metodologii upravleniya konkurentosposobnost'yu avtoproizvoditelya v usloviyah promyshlennoj sborki: dis. dokt. tekhn. nauk: 2.5.22 / Klentak Anna Sergeevna. – Samara, 2025.
4. *Klentak, A.S.* Koncepciya metodologii prognozirovaniya vostrebovannogo potrebitelyami kachestva produkcii mashinostroeniya pri vyhode kompanii-proizvoditelya na novye konkurentnye rynki / A.S. Klentak, V.N. Kozlovskij, A.V. Gusev // Izvestiya Samarskogo nauchnogo centra Rossijskoj akademii nauk. – 2024. – T. 26. – № 1(117). – S. 5-14.
5. *Robertson S., Spärck Jones K.* Relevance weighting of search terms // Journal of the American Society for Information Sciences. 1976. № 27(3). S. 129–146.
6. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proc. NAACL. 2019.
7. *Reimers N., Gurevych I.* Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proc. EMNLP. 2019.
8. *Cormack G.V., Clarke C.L.A., Buettcher S.* Reciprocal Rank Fusion outperforms Condorcet and individual rank learning methods // Proc. 32nd ACM SIGIR. 2009. S. 758–759.

Vladislav Piunov, Postgraduate Student. E-mail: Taimkraft@mail.ru,