Análise do uso de recuperação da informação em documentos de atendimento - Estudo de caso em bases de soluções de informática

Alternative Title: Analysis of Information Retrieval in Call Center Documents - Case Study in Computer Solutions Bases.

Leandro dos Santos Gaspar
Universidade Federal Fluminense
Rua Recife, s/n, Jd. Bela Vista
Rio das Ostras - RJ
lsg.gaspar@gmail.com

Flávia Cristina Bernardini
Laboratório de Inovação no Desenvolvimento de Sistemas - LabIDeS / Instituto de Ciência e Tecnologia - ICT / Universidade Federal Fluminense
Rua Recife, s/n, Jd. Bela Vista
Rio das Ostras - RJ
fcbernardini@id.uff.br

Leila Weitzel
Universidade Federal Fluminense
Rua Recife, s/n, Jd. Bela Vista
Rio das Ostras - RJ
leila.weitzel@id.uff.br

RESUMO

Centrais de Atendimento buscam ser mais produtivas realizando um atendimento padronizado para os seus clientes. A fim de alcançar este objetivo, são utilizados procedimentos, que contém um conjunto de soluções possíveis. O motor de busca atual usa um modelo booleano simplificado e não atende as necessidades de desempenho e consistência da Central de Atendimento. Esta pesquisa tem como objetivo descobrir qual método de recuperação de informação, por exemplo, vetorial ou probabilística, têm melhor desempenho em um motor de busca.

Palavras-Chave

Recuperação da Informação, Okapi BM25, Similaridade baseada em cosseno, Central de Atendimento

ABSTRACT

Call Centers aim to be more productive by performing a standard service for its customers. In order to accomplish this goal, it is used procedures, which contains a set of likely solutions. It must be stressed that the current engine uses a simplified Boolean model, and left the systems less consistent e slow with the Call Center needs. This research aims to figure out which information retrieval methods, e.g., vector or probabilistic, have better performance in a search engine.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Relevance Retrieval models, Search process, Selection process

General Terms

Algorithms, Experimentation Measurement

Keywords

Information Retrieval, Okapi BM25, Cosine Similarity Vector, Call Center

1. INTRODUÇÃO

Em uma Central de Atendimento (CA), para que se ocorra um atendimento padronizado e com maior produtividade no atendimento às solicitações dos clientes, é necessário que as informações estejam organizadas de modo apropriado. As informações devem estar armazenadas como roteiros, denominados *scripts* ou procedimentos de atendimento. Os procedimentos de atendimento são documentos eletrônicos que contêm orientação para executar tarefas relativas aos produtos e serviços da instituição. O objetivo desses procedimentos é resolver problemas, reclamações, dúvidas e elogios dos clientes [1].

Mesmo com a importância dos procedimentos de atendimento, um desafio atual é recuperar o(s) procedimento(s) mais relevantes para resolver um determinado problema do cliente. Com o aumento da capacidade de armazenamento dos computadores atuais, tornou-se relativamente barato armazenar grandes quantidades de dados em formato digital [2], facilitando o crescimento das bases de procedimentos. Porém, mesmo com uma base de procedimentos de atendimento rica em soluções de problemas, a recuperação da informação pode ser dificultada pelo formato do seu registro.

O cenário descrito anteriormente é observado na CA responsável por atender os chamados da força de trabalho de uma empresa relacionada ao ramo de petróleo situada no Brasil. A empresa possui muitos funcionários, a maioria deles utiliza computadores ligados em rede para executar suas atividades. São vários sistemas diferentes sendo utilizados dentro da empresa. Essa grande quantidade de pessoas, sistemas e equipamentos demandam muitos chamados, que podem ser simples dúvidas na operação de algum sistema até a correção no funcionamento de algum equipamento. Esse contexto justifica a necessidade de se gerenciar a manipulação desses procedimentos de forma que atenda às especificidades da CA.

Ainda, na mesma empresa devido ao grande número de sistemas e regiões atendidas, a possibilidade de dúvidas e solicitações dos clientes pode ser muito variada. Existem casos que, para a mesma solicitação, o atendimento é diferente, dependendo da região que é feita a solicitação. Para garantir que o atendente passe as informações corretamente, ele utiliza o "Portal do atendimento". Esse sistema faz uso de uma base de dados que contém todos os procedimentos de atendimento que detalham como funciona os sistemas, qual equipe atende que tipo de chamado, como instalar um determinado software, entre outras informações. O uso deste sistema para a busca do procedimento evita que o solicitante fique muito tempo à espera de um atendimento. Caso o chamado não possa ser resolvido de forma remota, o chamado é encaminhado para uma equipe especializada ou para uma equipe que preste o atendimento localmente ao usuário.

Para atender de forma satisfatória aos funcionários da empresa e possuir um canal único de contato, a área de atendimento ao usuário do setor de tecnologia da informação e telecomunicação decidiu contratar uma central de serviço para prestar o primeiro atendimento aos usuários. Os atendentes da CA prestam informações, atendem solicitações relacionadas à área de tecnologia de informação e telecomunicação, e encontram solução dos problemas no menor prazo possível.

Atualmente existem três formas do funcionário da empresa abrir um chamado, que são encaminhados para a CA: pelo telefone da central de serviço; através de um sistema na intranet; ou através de um correio eletrônico para o endereço da central de serviço. Por exemplo, no mês de setembro de 2014 foram registradas 159541 ligações atendidas, 22389 registros via web 12895 solicitações via nota de correio em uma das CA. Todo pedido de atendimento, independente do meio utilizado para a solicitação, deve ser registrado em um sistema de gestão de chamados, gerando um número único para cada atendimento. Cada registro de atendimento deve possuir, antes de ser finalizado: as informações descritas pelo usuário; a descrição das ações executadas pelo atendente para tentar solucionar o problema; a classificação do atendimento; o tipo de sistema ou equipamento que originou o chamado. Assim, ficam registrados: o tempo de atendimento; o contato da pessoa que abriu o chamado; os atendentes envolvidos; equipes que trabalharam para a solução do pedido.

Os tipos de pedidos de atendimentos podem ser classificados de modo genérico em: (i) **Solicitação -** Os chamados são classificados desta maneira quando se trata de um pedido de novos serviços ou dúvidas. Por exemplo, solicitação de instalação de um novo software em um Computador Pessoal (*Personal Computer* - PC), pedidos de acesso a algum sistema, entre outros. (ii) **Incidentes -** Os chamados são classificados desta maneira quando se trata de uma ocorrência (ou evento) não planejada ou inesperada. Por exemplo, um PC que não está ligando, um sistema que esteja “travando”, etc.

Existem atualmente duas CA’s em funcionamento na empresa no qual é realizado o estudo de caso. As centrais de serviços funcionam como contingência uma da outra, de forma que se alguma tiver algum problema no seu funcionamento as ligações são redirecionadas para a outra central de serviço.

2. TRABALHOS CORRELATOS

Nesta seção são apresentados os trabalhos que orientaram esta pesquisa. O objetivo em comum destas investigações é o uso de diferentes métodos de recuperação de informação em bases de dados de CA.

No trabalho realizado em [3], é proposta uma nova forma de organização dos documentos de um corpus para facilitar a sua recuperação. A pesquisa foi realizada na mesma empresa de petróleo utilizada nesta pesquisa, porém, na área de atendimentos diversos (serviços de telefonia, assistência de saúde, atendimento a fornecedores, entre outros). Porém, é utilizado algoritmo de clusterização e a criação de rótulos através de algoritmo de classificação para se propor uma nova forma de organização do corpus. Este trabalho serviu de inspiração, pois mostra que o problema de localização de procedimentos em centrais de atendimentos ocorre em várias áreas. Contudo, na pesquisa deste artigo, optou-se por utilizar métodos de similaridade para se localizar o documento em vez de se propor uma nova forma de organização do corpus.

O uso de recuperação da informação em uma base de atendimento também foi pesquisado em [4]. Nesse trabalho, os autores falam sobre uma base de dados de problemas que utilizava uma busca booleana. Com o crescimento da base de conhecimento, a busca passou a trazer muitos documentos irrelevantes para a consulta realizada. Segundo o artigo, o mais frustrante para os usuários dessa base era saber que o documento desejado estava armazenado e não conseguir encontra-lo. O problema enfrentado nesse trabalho é muito parecido com o que se deseja resolver nesta pesquisa, pois o sistema atual realiza busca apenas no título do documento e também ocorre o problema de alguns procedimentos não serem encontrados pela consulta, apesar de se saber que eles estão registrados. O foco do trabalho realizado em [4] foi melhorar a precisão da busca, para isso eles implementaram uma forma de transformação visual e interativa no texto da busca. Com essa mudança, o usuário passou a ver quantos artigos possuem o termo informado, pode trocar o termo por algum sinônimo sugerido pelo sistema e mudar os operadores lógicos utilizados na consulta. Diferentemente do trabalho realizado em [4], o foco da pesquisa deste artigo é realizar a indexação de todo o conteúdo do documento e avaliar os resultados utilizando o método de busca CVS (Cosine Similarity Vector) e BM25 (BM stands for Best Matching) para atribuir pesos aos termos dos documentos a fim de se obter uma busca mais precisa e ordenada.

No contexto de avaliação de algoritmos, em [2] foi realizada uma comparação entre os métodos de recuperação de informação BM25, TF-IDF (Term Frequency–Inverse Document Frequency**)** e CVS. Nesse trabalho, o autor utiliza um documento como referência, para extrair expressões multipalavras (EM) e buscar documentos similares. Para a realização da avaliação, o autor desenvolveu dois componentes de um sistema, um denominado *server*, responsável por indexar os documentos do corpus e outro denominado *client*, responsável por receber o documento de referência para a pesquisa, extrair as expressões multipalavras e realizar a busca. Para realizar a avaliação entre os métodos foram indexados 184 documentos, depois foram utilizados dez documentos diferentes dos que foram indexados e realizada a busca. Para cada um dos documentos, foram realizadas três buscas, uma para cada um dos métodos propostas, e verificado a quantidade de documentos retornados. Sua avaliação final é que o método BM25 é o método que retornou mais documentos e que o método CVS é o mais seletivo. Os autores concluem que o método BM25 é o método mais apropriado, pois leva em consideração os diferentes tamanhos de documentos e o peso das expressões multipalavras. Esse trabalho possui grande correlação com o que se busca fazer neste trabalho, pois se deseja avaliar

qual é a melhor metodologia de indexação e busca para uma base de procedimentos.

Diversos outros trabalhos, como p. ex. [5,6,7], utilizam métodos de recuperação da informação para resolver diversos problemas distintos. O que pode ser observado nesses trabalhos é que a questão de qual o melhor método de recuperação de informação depende dos parâmetros utilizados para avaliação e da base de dados avaliada. Deste modo, um método de recuperação da informação pode apresentar melhor resultado em uma base e não apresentar resultados tão bons em outra base de dados. Por esse motivo, este trabalho irá avaliar qual dos métodos, (BM25 ou CVS) traz melhor resultado na base estudada.

3. METODOLOGIA

A metodologia adotada neste trabalho envolve quatro etapas:

Etapa1 – Definição da base de dados para o experimento e definição do tamanho da amostra;

Etapa2 – Fase de pré-processamento que envolve:

(i) **Análise Léxica** (conhecida como *tokenização*) para a retirada de caracteres especiais tais como /, %, $ e etc. São aceitas apenas as letras de a-z, os números 0-9 e os símbolos @ (arroba), # (tralha) e o apóstrofe. Também tem como objetivo decompor o texto em unidades estruturais menores, em nosso caso as unidades estruturais são as palavras.

(ii) **Remoção das *stopword***, **s**egundo [8], uma *stopword* pode ser traduzida "palavra vazia", elas aparecem em praticamente todos os documentos, ou na maioria deles, por isso não são capazes de colaborar na análise da polaridade do texto.

(iii) **Stemização,** que reduz palavras distintas a sua raiz gramatical comum;

Etapa3 - Indexação dos termos;

Etapa4 – Recuperação de documentos. Podem ser utilizadas diferentes funções de similaridade para recuperação de documentos, que considera como parâmetro o texto de busca realizada b e os documentos indexados na base de dados - D = {d1, d2, ..., dN}. Duas funções comumente usadas são a CVS (Cossine Vector Similarity) e a Okapi BM25 (neste trabalho, tal medida é referenciada somente como BM25). A primeira - CVS - é uma medida de similaridade entre dois vetores de um espaço com produto interno que mede o cosseno do ângulo entre esses vetores (quanto maior o valor do cosseno, maior a similaridade entre os dois vetores); a segunda - BM25 - é uma função de ordenamento baseada em um framework de recuperação probabilístico [9].

4. ESTUDO DE CASO E RESULTADOS

Conforme definido na Etapa 1 da metodologia, foi utilizada a base de procedimentos da CA de informática de uma empresa de petróleo no experimento. Para calcular o número de textos *n* na amostra, foi utilizada a equação da curva de Gauss [10] conforme apresentado na Equação 1. Nessa equação, *n* é o tamanho da amostra a ser calculada; *N* é o tamanho da população (no caso do experimento, 133.115 documentos em formato texto); *Z* é a variável normal padronizada associada ao nível de confiança (no estudo de caso, foi considerado um nível de confiança de 95%); *p* é a probabilidade do evento (no caso, $p = 0{,}5$); e *e* é o erro amostral (no caso, $e = 0{,}05$). Daí, para a condução do estudo de caso, foi retirada aleatoriamente uma amostra de 374 documentos.

$$n = \frac{N.Z^2.p.(n-p)}{Z^2.p.(1-p) + e^2.(N-1)} \quad (1)$$

Após definido o tamanho da amostra, foram extraídos documentos aleatoriamente da base e adicionados ao sistema de arquivos. Na execução da Etapa 2, verificou-se então que os documentos foram criados a partir de um modelo que possuíam várias informações desnecessárias para o objetivo deste trabalho. Exemplos destas informações são: quem criou o documento, quem aprovou, sigla da gerência responsável pelo documento, data da criação. Estão informações eram preenchidas em formato de tabela diretamente no documento. Antes da indexação, precisou-se então passar por uma fase de limpeza dos documentos, onde se retiraram todas estas informações desnecessárias e gravou-se o texto resultante em formato de texto padrão ANSI.

Após a transformação dos documentos, foi desenvolvido um protótipo que faz a leitura desses textos da amostra, realiza a análise léxica, sintática, remoção das stopwords e stemização dos termos. O protótipo também executa a Etapa3 realizando a indexação, utilizando TFIDF como peso para os termos. Após a indexação, o protótipo permite digitar uma expressão para a busca dos documentos. Após realizar o pré-processamento e cálculo do peso dos termos da consulta foi realizado o cálculo de similaridade com os termos da base indexada. A Tabela 1 apresenta o resultado de uma consulta realizada no protótipo, listando os 10 primeiros procedimentos retornados por cada método.

CONSULTA: *"Como solicitar restore de arquivameto"*	
Resultado BM25	**Resultado CVS**
1. Documento A	1. Documento F
2. Documento B	2. Documento K
3. Documento C	3. Documento D
4. Documento D	4. Documento G
5. Documento E	5. Documento H
6. Documento F	6. Documento I
7. Documento G	7. Documento J
8. Documento H	8. Documento L
9. Documento I	9. Documento M
10. Documento J	10. Documento N

Tabela 1 - Tela de consulta de procedimento

Após a realização de dez consultas para cada método avaliado, verificou-se, por validação manual, que os resultados se mantiveram constantes, quanto a quantidade de documentos relevantes retornados por cada método.

Especificamente, na consulta apresentada na Tabela 1, foram retornados seis documentos relevantes pelo método BM25 nas posições 1,2,3,4,6,8 e quatro documentos relevantes pelo métoo CVS nas posições 1,3,5,8. Também verificou-se que os documentos retornados nas posições dois e três pelo método BM25 não foram retornados no método CVS. A Tabela 2 apresenta a correspondência de posição dos documentos

relevantes retornados por cada método. Deve ser observado que, para o método CVS, o símbolo "-" significa que os documentos recuperados nas posições dois e três não são relevantes para a busca. Pode ser observado nessa tabela que o método BM25 conseguiu recuperar cinco documentos relevantes em seis, e ainda com correspondência para os três primeiros documentos relevantes. Já o CVS recuperou somente três dos primeiros seis documentos relevantes, sem correspondência de posição.

Documentos relevantes	Posição no método BM25	Posição no método CVS
1	1	8
2	2	-
3	3	-
4	6	1
5	4	3
6	8	5

Tabela 2 – Correspondência dos documentos relevantes

5. CONCLUSÕES E TRABALHOS FUTUROS

É de extrema importância que, no desenvolvimento de Sistemas de Informações que dão suporte a Centrais de Atendimento (CAs), o processo de recuperação de soluções de problemas seja realizado de maneira efetiva, que realmente retorne resultados relevantes para os atendentes e/ou usuários. No entanto, são diversos os casos nos quais esse é um problema latente, como é o caso de uma CA responsável por atender os chamados da força de trabalho de uma empresa relacionada ao ramo de petróleo situada no Brasil. Assim, neste trabalho é explorado o uso de técnicas de recuperação da informação para melhorar a recuperação de documentos. Os resultados dos experimentos mostraram que os métodos testados melhoram o retorno da busca em relação ao método atualmente utilizado. A melhoria dos resultados está relacionada aos pesos dos termos nos documentos. No modelo que vem sendo utilizado, por várias vezes os documentos não eram retornados na consulta.

Os resultados apresentados na seção 4 sugerem que o método BM25 tem um melhor desempenho no retorno das consultas. Porém, para validarmos esta proposta, é necessário que se tenha uma lista de documentos relevantes para cada situação no CA. Sendo assim, como trabalho futuro pretende-se utilizar a técnica de "crowdsourcing" para rotular os documentos em função da sua relevância em determinado tema, essa base anotada servirá para validar os resultados obtidos nos testes iniciais.

6. REFERÊNCIAS

[1] SILVEIRA, Sandra Maria; MOURA, Maria Aparecida. Scripts de atendimento em call centers: uma visão de documentos eletrônicos. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 15, n. 29, p. 145-168, 2010.

[2] SOUZA, Renato Rocha. Comparing three different techniques to retrieve documents using multiwords expressions.

[3] DA SILVA, Edson Marchetti; SOUZA, Renato Rocha. Comparing three different techniques to retrieve documents using multiwords expressions. In:**10 CONTECSI**. 2013.

[4] ANICK, Peter G. Integrating natural language processing and information retrieval in a troubleshooting help desk. **IEEE expert**, v. 8, n. 6, p. 9-17, 1993.

[5] XU, Lixin; CHEN, Guang; YANG, Lei. Incremental clustering in short text streams based on BM25. In: **Cloud Computing and Intelligence Systems (CCIS), 2014 IEEE 3rd International Conference on**. IEEE, 2014. p. 8-12.

[6] ESTEVA, Maria; BI, Hai. Inferring intra-organizational collaboration from cosine similarity distributions in text documents. In: **Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries**. ACM, 2009. p. 385-386.

[7] BIGDELI, Elnaz; BAHMANI, Zeinab. Comparing accuracy of cosine-based similarity and correlation-based similarity algorithms in tourism recommender systems. In: **Management of Innovation and Technology, 2008. ICMIT 2008. 4th IEEE International Conference on**. IEEE, 2008. p. 469-474.

[8] WIVES, Leandro K.; LOH, Stanley. Recuperação de informações usando a expansão semântica e a lógica difusa. In: **Congreso Internacional En Ingenieria Informatica, ICIE**. 1998.

[9] BAEZA-YATES, Ricardo et al. **Modern information retrieval**. New York: ACM press, 1999.

[10] PASQUALI, Luiz. A Curva Normal. 2006.