UM BUSCADOR PARA GRANDES ACERVOS: um estudo de caso com o jornal a Tribuna no Espírito Santo

Abeil Coelho Júnior[1]
Elias de Oliveira[2]

RESUMO: A despeito dos avanços já vivenciados com buscadores tais como Google, Bing, Yahoo e muitos outros, ainda carecemos de boas soluções para lidarmos com documentos arquivísticos em acervos locais. Esse artigo discute algumas das características de uma solução desenvolvida pelo grupo de Recuperação Inteligente da Informação da Universidade Federal do Espírito Santo, quando provendo acesso às imagens digitalizadas de um acervo de mais de 200.000 páginas. Atualmente, discutimos também como essa ferramenta está permitindo o acesso e também a mineração de dados no acervo testado. Comparamos essas facilidades de acesso àqueles oferecidos pela Hemeroteca Digital da Biblioteca Nacional.

Palavras-chaves: Buscador. Jornal. Mineração de dados.

A SERCHER FOR LARGE COLLECTIONS: a study of case with the newspaper a Tribuna in Espírito Santo

ABSTRACT: In spite of the progress we are living with search engines such as Google, Bing, Yahoo and others, we still lack of good solutions to deal with access in local archives collections. This article discusses some of characteristics of a solution developed by the group of intelligent recovery of information at Federal University of Espírito Santo, when giving access to digitalized images from a collection of more than 200.000 pages. Nowadays, we discuss how our tool is allowing the access and the performing datamining into our large collection. We compared these facilities of access with the offered by the Digital Newspaper Library of the National Library.

Keywords: Search Engine. Newspaper. Data Mining.

1 INTRODUÇÃO

Um dos papéis do arquivista é dar acesso às informações que esse profissional seja curador (BELLOTTO, 2003). Visto que o arquivista no Século XXI se encontra na era da informação, aonde se cria tanta informação que não temos como dar conta de consumi-la. É, portanto, fundamental o uso e o desenvolvimento de novas ferramentas eletrônicas e eficientes para a administração dessas informações. Com isso o papel do arquivista se desloca, de uma postura tradicional, quando da primazia do uso do papel, para uma melhor capacitação e entendimento das novas tecnologias, suas vantagens e riscos sobre o suporte anterior. Somente assim é que esse profissional estará capacitado a dar acesso rápido aos objetos

[1]Estudante de Arquivologia na Universidade Federal do Espírito Santo (UFES). abeilc@hotmail.com
[2]Professor na Universidade Federal do Espírito Santo (UFES). elias_de_oliveira@yahoo.com.br

documentais nos novos suportes de informações de maneira rápida e precisa (BELLOTTO, 2003, BOERES, 2006, SAYÃO E SALES, 2012).

O aumento do volume de informação desagregadas e, portanto, fora de uma organicidade de interesse arquivístico, também tem sido um desafio para os profissionais de tecnologia (JAGADISH, ET AL. 2014). O que só deve aumentar nossa preocupação com respeito aos efeitos desse problema para os futuros profissionais da área da ciência da informação.

O grupo de pesquisa de Recuperação Inteligente da Informação, no Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo, vêm desenvolvendo vários algoritmos de mineração de dados e textos para várias aplicações (OLIVEIRA ET AL., 2015, SAÚDE ET AL., 2014). Dentre essas aplicações, o aLine é um sistema complexo com várias funcionalidades. Uma das funcionalidades que queremos discutir nesse trabalho é a de servir como máquina de busca para o acervo do Jornal A Tribuna, um jornal popular local no Espírito Santo. Hoje já indexamos um acervo com mais de 200.000 imagens de páginas digitalizadas desse jornal.

Esse projeto mira-se naquilo que foi dito pelo arquivista canadense (COOK, 2001) quando afirma que, os meios analógicos, com o passar do tempo, irão se tornar quase totalmente irrelevantes. Esse autor acredita que o destino do meio físico é de se deteriorar e, antes que isso aconteça, devemos fazer a migração para o meio digital. Então, assim nos parece, nosso futuro será digital.

O aLine é uma ferramenta-laboratório para a mineração de dados sobre esse acervo, e em breve outros acervos, que além de trazer resultados de quaisquer termos existentes nos documentos indexados também gera gráficos da evolução quantitativa do uso dos termos na base, ao longo do tempo, com apenas alguns toques. Em suma, hoje nos é claro que utilizar buscadores inteligentes para auxiliar pesquisas em acervos, em uma ou várias bases grandes tais como essa do jornal A Tribuna, economiza muito tempo do usuário pesquisador. As unidades de informação não podem ficar à margem dessa nova tecnologia, dessa nova ferramenta de trabalho.

Este artigo está organizado da seguinte forma. Na próxima seção, descrevemos o acervo com o qual estamos atualmente trabalhando e do qual extrairemos alguns exemplos de resultados. Os trabalhos relacionados serão discutidos na Seção 3. O buscador aLine é descrito na Seção 4, quando apresentamos algumas características quantitativas de desempenho e algumas funcionalidades já disponíveis. Realizamos algumas comparações e os resultados discutimos na Seção 5. Finalizando, na Seção 6, apresentamos nossas conclusões e

trazemos à luz algumas propostas que entendemos ser interessantes como continuação desse projeto.

2 O ACERVO

Figura 1: Exemplo de uma página do caderno AT2

PAULO OCTAVIO

Vitrine

Dicas para o verão

Bronze saudável

Sono garantido

Dieta para dormir

Doce refresco

Uniforme

Abraços partidos

Quem disse

Você sabia?

.

O aLine indexa todos os termos/palavras dos documentos incorporados aos acervos, ao converter os documentos vindos de formatos *pdfs*, ou outros formatos, para arquivos de texto. Optamos por esse método, porque queremos que a indexação seja automática e que as buscas possam ser realizadas sobre qualquer termo que leve a um documento da base.

Hoje, o acervo do aLine possui aproximadamente 200 mil documentos, imagens de páginas que datam de 2003 a 2010 do jornal A Tribuna. Esse jornal edita, periodicamente os seguintes cadernos listados na Tabela 2:

Tabela 1: Cadernos do jornal A Tribuna.

Classes		
AT2	Noticiário	Sobre Rodas
Jornal da Família	Tv Tudo	Mulher
Minha Casa	Informática	Especial
Imóveis	Classificados	Tudo a Ver

Cada um desses cadernos pode ter de uma a mais de dez páginas. Quase todos os títulos dos cadernos são quase autoexplicativos, por outro lado, o caderno AT2 é um caderno de variedades. Apresentamos um exemplo de página desse caderno na Figura 1.

3 TRABALHOS RELACIONADOS

A comunidade arquivística brasileiras buscam arduamente encontrar caminhos para fazer frente ao desafio da grande massa documental a ser gerida e preservada nos dias de hoje. Entre muitas outras estratégias, vemos o esforço por se encontrar ferramentas automatizadas que possam ajudar no processo metodológico de solução desse problema.

Flores (FLORES; HEDLUND, 2014) é um dos autores que em muito vem contribuindo para a formação dos novos profissionais, treinamento continuado de outros e, também, tornar uma *praxis* comum a utilização de algumas ferramentas, já bem consolidadas em experimentos práticos. No artigo (FLORES E HEDLUND, 2014), o autor discute um desses resultados práticos oriundo de um trabalho em nível de mestrado, onde discute e avalia a utilização do *software* ICA-AtoM na atividade de descrição arquivística e o acesso por meio da internet por parte de muitos usuários. Nesse trabalho, o autor realiza seus experimentos com um acervo fotográfico do Arquivo Histórico Municipal de Santa Maria (AHMSM), RS. Descreve que esse acervo estaria, originalmente, em suporte papel e constituem um universo de 733 unidades documentais, divididas e armazenadas em 17 pastas-arquivo e em quadros emoldurados.

Uma das características importantes ao se avaliar uma ferramenta de *software* é se medir a capacidade da mesma em dar conta de garantir bom desempenho de forma escalável. No caso do trabalho, citado anteriormente, a quantidade de objetos registrados no banco de dados não seria o suficiente para se avaliar essa característica.

Em (GUEGUEN, ET AL, 2013), um trabalho é publicado como resultado de um levantamento realizado pelo Grupo de Especialistas em Descrição Arquivística (GEDA) no final de 2012 (GUEGUEN, ET AL. 2013). Esse grupo teve como desafio desenvolver um modelo conceitual para descrição arquivística. Esses modelos tinham que conciliar e integrar as quatro normas de descrição do Conselho Internacional de Arquivos (CIA) já existentes.

Como parte do processo de trabalho do GEDA, eles avaliaram o ICA-AtoM. Na avaliação, observaram que o conceitual implementado nesse *software* exibe proeminentemente materiais de arquivo (documentos), agentes, entidade custodiadora e eventos, entre outros aspectos de maior interesse para áreas como, por exemplo, a comunidade das bibliotecas.

Uma ferramenta de apoio e facilitação não é tudo, não resolve todo o trabalho que é demandado ao especialista da informação. Por isso, Humberto Innarelli diz que:

> Preservar o passado para conhecer, aprender e inovar é fundamental não só como vantagem competitiva, mas também, como forma de desenvolvimento e manutenção da Cultura de nossa Sociedade (..)

O autor (INNARELLI, 2012) busca estabelecer um vínculo entre o papel da gestão arquivística, em particular quando se referindo aos documentos digitais e a gestão do conhecimento explícito registrados nos novos suportes digitais. Salienta que, a despeito do esforço de investimento em Gestão do Conhecimento e Tecnologias da Informação e Comunicação, há uma negligência de conceitos já sedimentados em áreas tais como a Arquivística, em prol da rentabilidade e a competitividade a todo custo entre as instituições.

A seguir, descrevemos brevemente o *software*, o qual é hoje considerado uma referência na área como aquele que atende, se não todos, quase todas as sugestões de padrão sugerido pelo ICA (GUEGUEN, ET AL., 2013).

4 O BUSCADOR ALINE

A máquina de busca aLine foi, inicialmente, concebida para coletar e indexar páginas do jornal A Tribuna no Espírito Santo. A ideia é termos um buscador, para o acervo do jornal, semelhantes aqueles de propósito geral que já conhecemos: Google, Yahoo! e Bing. Esses últimos têm como objetivos a indexação de páginas *Webs*. Por isso, qualquer base ou documento que não esteja na *Web*, estarão fora do alcance deles. Todavia, acreditamos haver

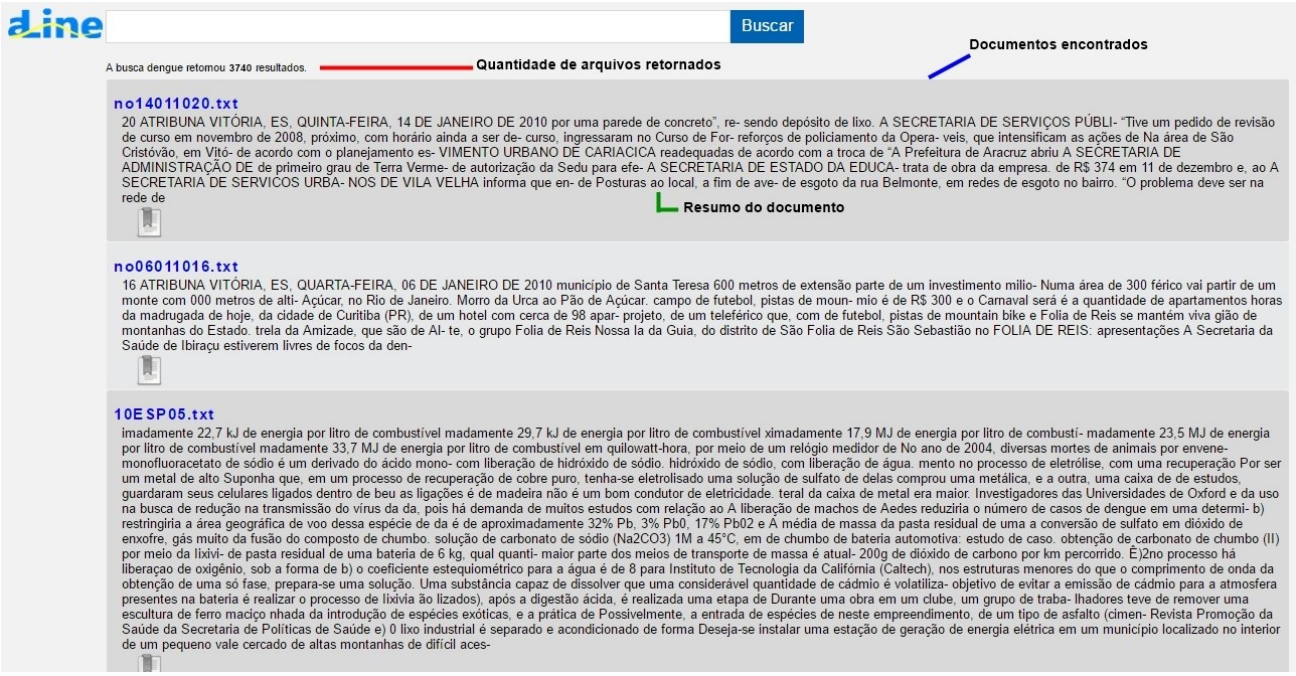
muito mais coleções de documentos nas intranets do que dados formatados na *deep web* (MADHAVAN, ET AL., 2008).

O aLine possui hoje duas páginas básicas: uma página HTML inicial e a de resultados. A Figura 2 mostra a página inicial do buscador aLine, com uma caixa de busca e um botão para submeter a busca. Nesse exemplo, foi realizada uma busca do termo *dengue*.

Figura 2: Recorte da página inicial.

Uma vez tendo executado a busca, a Figura 3 mostra a página de parte dos resultados do buscador aLine sobre o acervo existente até o momento da escrita desse artigo. No topo da página, vemos novamente a caixa de busca. Logo abaixo, a quantidade de documentos retornados e a lista com os *links* dos resultados da busca, seguidos dos resumos das páginas, respectivamente. Uma caixa de busca foi mantida nessa segunda página, para que o usuário possa fazer outras buscas sem a necessidade de voltar para a página inicial. Através dos *links* o usuário é conduzido à página digitalizada do jornal, a qual contém os termos pesquisados.

Figura 3: Página de resultados.



O resumo na página de resultados corresponde a tentativa de construirmos uma síntese do que há na página encontrada. As pesquisas nessa funcionalidade ainda são muito incipientes no aLine. Aqui vemos um grande desafio de pesquisa tendo em vista a especificidade de uma página de jornal a qual, na mesma página, discorre-se sobre vários assuntos ao mesmo tempo.

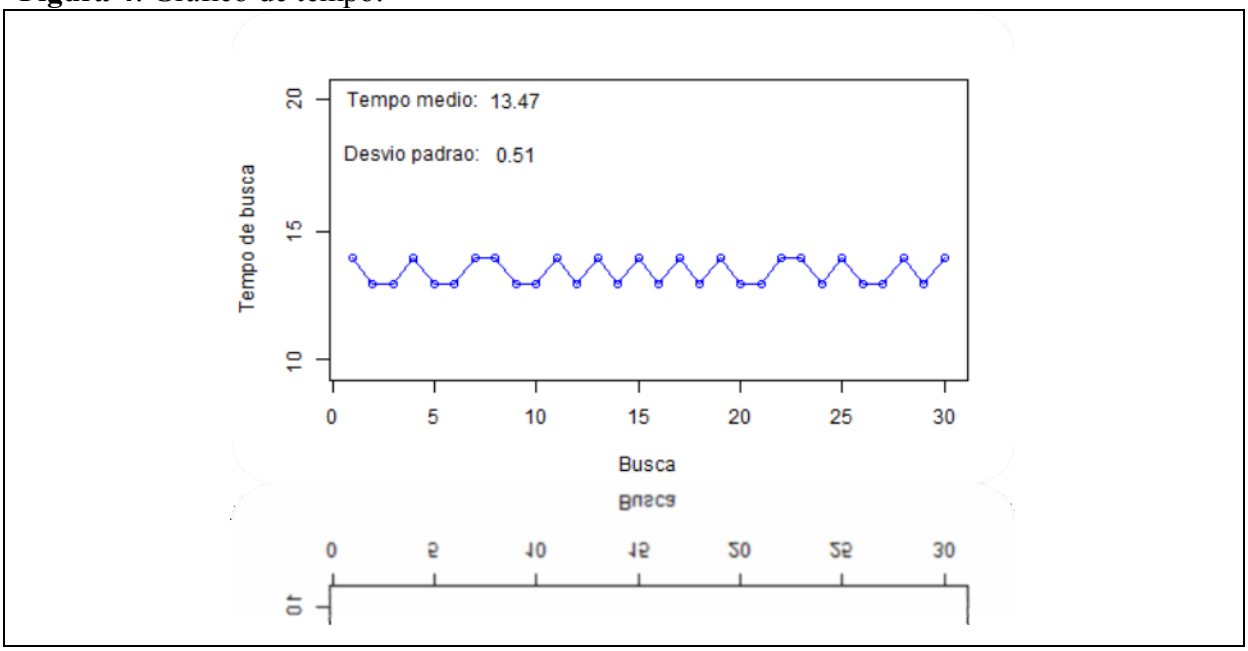
Além do desafio metodológico de se propor uma melhor estratégia de conceituar o que seria um resumo, ou síntese, de uma página de jornal, nos deparamos também com outro desafio de como apresentar esse resultado nas diversas mídias populares hoje em dia, tal como o celular. Até o momento da escrita desse artigo, a decisão de apresentação dos resultados, quando apresentados em um aparelho móvel, é de omiti-los. Apresentamos apenas os *links* que levarão às páginas que contenham os termos pesquisados.

4.1 ALGUMAS MEDIDAS DE DESEMPENHO

Fizemos alguns experimentos para avaliarmos o desempenho de nosso buscador. Nesses experimentos construímos um programa (*script*) para sortear uma palavra e despachar uma busca no aLine. Fizemos isso com mais de 1000 termos individuais e combinados de 2-5 termos por vez. Cada uma busca específica foi repetida várias vezes para que pudéssemos medir a variância de tempo de resposta e, assim, calcular algumas estatísticas.

Em um dos experimentos, utilizamos a palavra *Presidente* numa amostra de 45 mil documentos do acervo. Realizamos 30 repetições da busca, a qual retornou sempre a quantidade de 6121 documentos, ou seja, aqueles documentos que possuem a palavra *Presidente*. O tempo médio de busca foi de 13.47 segundos com o desvio padrão de 0.51. Como mostra a Figura 4.

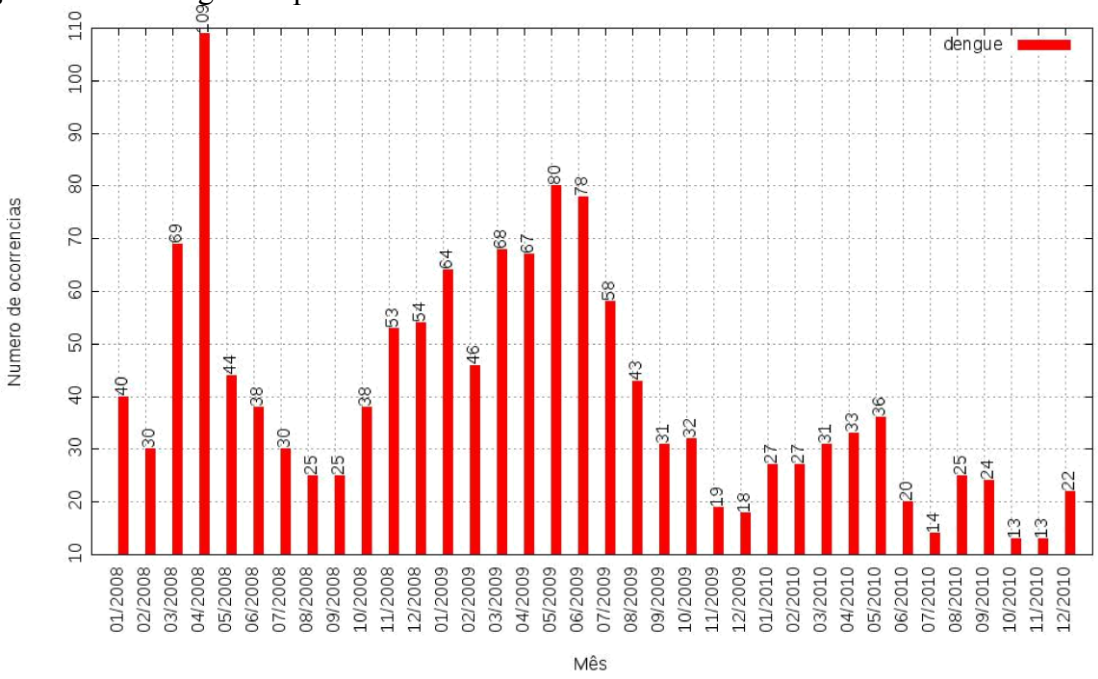
Figura 4: Gráfico de tempo.



4.2 ALGUNS OUTROS SERVIÇOS

Com vista a atender algumas demandas de pesquisadores colegas, logo vimos a oportunidade de dar provimento, via o próprio sistema de busca aLine, de outros serviços relacionados a busca de documentos. Um desses que discutiremos brevemente aqui é aquele que denominamos de **report:**. Trata-se de um serviço a ser executado de forma assíncrona. Ou seja, além da busca com o retorno das respectivas páginas que contenham os termos como palavras-chave, no aLine o usuário tem funcionalidades que o permitem, ao fazer uma busca, obter de forma automática um relatório com gráficos para análises de tendências. A Figura 5 mostra o resultado de uma busca de documentos contendo o termo *dengue* como chave de busca.

Figura 5: Gráfico gerado pelo buscado



Essa funcionalidade é explorada no aLine via a utilização de um dos parâmetros de busca, o **report:**, nesse caso. O usuário, na barra de busca, além de colocar seus termos de busca, como anteriormente, agora ao acrescentar o parâmetro **report:**, na seguinte forma: (dengue report:meu-email@servidor.com) (veja também a Figura 6), estará dizendo ao buscador que quer o resultado enviado para o seu e-mail particular (meu-email@servidor.com). Por ser um serviço processado assincronamente no servidor do aLine, portanto não imediatamente, nos abre a possibilidade de acrescentar mais informações de interesse para o usuário, como por exemplo, a frequência dos termos de busca ao longo de um tempo. No caso em tela, esse o período considerado foi a totalidade do acervo. Para restringir a busca a um período temporal basta acrescentar mais alguns parâmetros. Por exemplo, a busca do mesmo termo anterior, *dengue*, apenas dentro no mês Abril de 2006, é feito da seguinte forma:

(dengue datefrom:20060501 dateto:20060530 report:meu-email@servidor.com)

Essa busca mudaria o formato do gráfico na Figura 6 para mostrar apenas a frequência dos termos pesquisados nesse intervalo.

Figura 6: Busca com o filtro **report:**.

4.3 UMA VISÃO DE FUTURO PRÓXIMO

O buscador aLine ainda está em desenvolvimento. Nós queremos aumentar a atual base, das edições do jornal A Tribuna de 2010 para até os dias atuais, e mantê-la atualizada a partir daí, conforme a publicação diária do jornal. Estimamos que, ao completar as edições apenas desse jornal, teremos em nosso acervo mais de 400 mil documentos – páginas do jornal. Hoje limitamos em 30 segundos o tempo de resposta para qualquer busca feita na página principal do aLine. Quando utilizando o **report:**, não temos limite de tempo. Nossa meta é, além de tentar diminuir ainda mais esse tempo limite para a resposta de buscas, é indexar outras bases de outros jornais, outros acervos, sem que esse tempo seja afetado. Esperamos em breve sermos capazes de lidar com federação de acervos de domínios diferenciados de forma individualizada, mas também, se for o caso, de forma unificada.

5 COMPARANDO COM UM SISTEMA REAL

Na Seção 3 discutimos alguns trabalhos os quais apontaram que um dos *softwares* candidato a se tornar padrão é o ICA-AtoM. Portanto, nós também entendemos que, esse *software* seja utilizado como plataforma básica tendo em vista que esse contempla os padrões acordados pela área arquivística como sendo os adequados para gerir e que possua os mecanismos necessários para melhor preservar esses documentos digitais para o futuro.

Contudo, mesmo o ICA-AtoM se utiliza de soluções de terceiros para implementar seu módulo de busca de forma eficiente. O aLine é concorrente nesse aspecto, podendo, portanto, estar dentro do ICA-AtoM como funcionalidade de busca. Assim, todas as funcionalidades descritas na Seção 4, poderiam ser incorporadas também ao ICA-AtoM.

Nós entendemos que os buscadores existentes, ou utilizados, pelos sistemas de gerência nas unidades de informação, são pobres de recursos gráficos e de análises inteligentes. Daí nosso interesse e nossas pesquisas no aprimoramento do ALINE e incremento de mais funcionalidades.

Para ilustrar nossa discussão, faremos uma comparação entre o que já dissemos na Seção 4 e aquilo que existe como recurso de busca em um sistema real. A Figura 7 é a imagem da página principal da Hemeroteca da Biblioteca Nacional, na época da escrita desse artigo. No site da Hemeroteca Digital, vemos alguns campos para filtrarmos nossa busca. Ao invés da utilização de parâmetros, como no caso do aLine, lá é utilizado vários campos para cada entidade de indexação.

O sistema de busca está organizado em três visões, as quais são implementadas através de três abas: periódico, período e Local. Essas abas servem para efetuar filtros para as buscas.

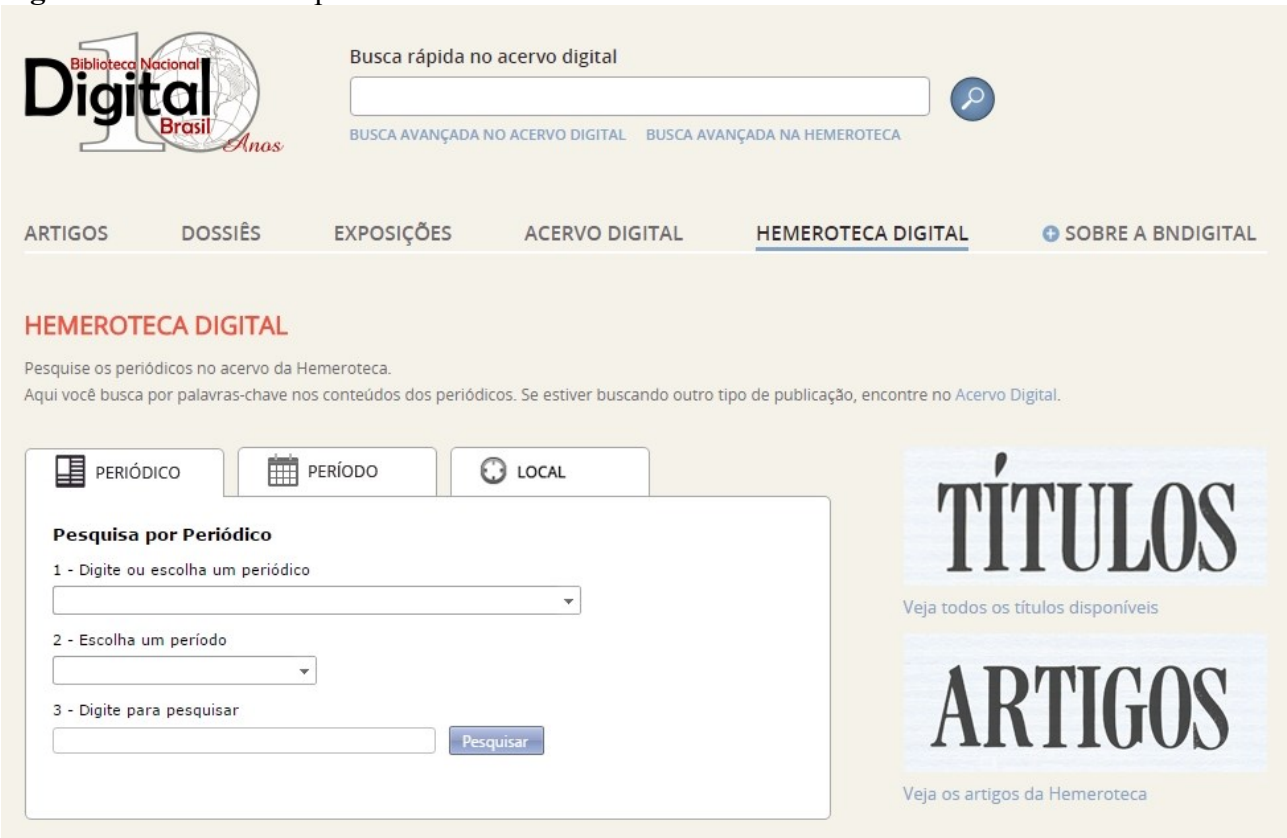
Nesse formato usuário deve preencher escolher um desses filtros primeiro para comandar sua busca. Por exemplo, se escolher fazer a busca com o foco primeiro em um acervo/periódico, ele deverá introduzir logo de início o nome do periódico. Ao escolher um periódico, uma lista para a escolha de períodos exclusivo para esse acervo é apresentada. Só depois disso é que o usuário será permitido introduzir seus termos de busca.

Por sua vez, o usuário que optar por dar foco no período primeiro, utilizará a aba de período e lá há um campo pré-formatado com períodos existentes correspondentes a todos os acervos. Só depois disso é que o usuário conseguirá passar para preencher os demais campos.

No caso do periódico possuir mais de uma ocorrência do termo buscado, navegamos para as próximas aparições pelas ocorrências através do uso da barra de navegação no topo da página de resultados. Esse periódico irá sendo explorado de acordo com as suas edições, sendo a ordem da versão mais antiga para as mais atuais.

Não identificamos no sistema a possibilidade de fazermos buscas com um conjunto de termos simultaneamente, de tal forma que um conjunto de termos possa aparecer na mesma página do periódico, por exemplo.

Figura 7: Recursos comparados.



6 CONCLUSÕES

Nesse artigo, descrevemos algumas das características técnicas de uma máquina de busca, ainda em desenvolvimento pelo grupo de pesquisa de Recuperação Inteligente da Informação. Esse buscador é capaz de realizar uma busca de várias palavras-chave, entre mais de 200.000 páginas, em menos de 30 segundos em média. As páginas localizadas são exibidas via o *click* do *link* dentre os resultados.

Para muito além do simples acesso, a ferramenta também provê serviços de mineração de dados capaz de processar a evolução cronológica de termos. Esse serviço nos permitiu analisar o crescimento do uso do termo Facebook e, respectivamente, a queda de uso do termo Orkut ao longo dos anos, desde 2007. Isso com apenas um *click*.

Comparamos os serviços oferecidos por essa ferramenta e aqueles existentes na Hemeroteca Digital da Biblioteca Nacional. Constatamos que, segundo algumas métricas estabelecida para as comparações, a máquina de busca aLine foi superior em todas as métricas aquela oferecida pela Hemeroteca da Biblioteca Nacional.

Como trabalho futuro, pensamos incluir mais funcionalidades facilitadoras para buscas mais inteligentes. Uma das próximas funcionalidades será a possibilidade de

especificarmos buscas especificamente sobre nomes de pessoas (CAMPOS E OLIVEIRA, 2015), ruas ou instituições. Finalmente, incluir esse buscador como de busca para o ICA-AtoM (AtoM) e, assim, poder gerenciar um conjunto de acervos de jornais de forma mais eficiente.

REFERENCIAS

BELLOTTO, H. L. *O Arquivista na Sociedade Contemporânea*. Marília, SC: Online, 2003. https://www.marilia.unesp.br/Home/Extensao/CEDHUM/texto01.pdf.

BOERES, S. Necessidade de Capacitação de Gestores para Preservação Digital na Biblioteconomia, Museologia e Arquivologia. *Revista Ibero-Americana de Ciência da Informação*, Revista Ibero-Americana de Ciência da Informação, Brasília, DF, v. 9, n. 2, jul. 2006.

CAMPOS, J.; OLIVEIRA, E. Extração de Nomes de Pessoas em Textos em Português: uma Abordagem Usando Gramáticas Locais. In: *Computer on the Beach 2015*. Florianopolis, SC: SBC, 2015.

COOK, T. Archival Science and Postmodernism: New Formulations for Old Concepts. *Archival Science*, New York, USA, v. 1, p. 3–24, 2001.

FLORES, D.; HEDLUND, D. Análise e Aplicação do ICA-AtoM como Ferramenta para Descrição e Acesso ao Patrimônio Documental e Histórico do município de Santa Maria – RS. *Informação & Informação*, João Pessoal, PB, v. 19, n. 3, p. 86–106, 2014.

GUEGUEN, G. et al. Para um Modelo Conceitual Internacional de Descrição Arquivística. *Revista Acervo*, Rio de Janeiro, RJ, v. 26, n. 2, 2013.

INNARELLI, H. Preservação Digital: a Gestão e a Preservação do Conhecimento explícito Digital em Instituições Arquivísticas. *InCID: Revista de Ciência da Informação e Documentação*, Brasília, DF, v. 3, n. 2, p. 48–63, 2012.

JAGADISH, H. V. et al. Big Data and Its Technical Challenges. *Commun. ACM*, ACM, New York, NY, USA, v. 57, n. 7, p. 86–94, jul. 2014.

MADHAVAN, J. et al. Google's Deep Web Crawl. *Proc. VLDB Endow.*, VLDB Endowment, v. 1, n. 2, p. 1241–1252, ago. 2008.

OLIVEIRA, E. et al. Using the Cluster-Based Tree Structure of k-Nearest Neighbor to Reduce the Effort Required to Classify Unlabeled Large Datasets. In: *7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Lisbon, Portugal: IC3K, 2015.

SAÚDE, M. R. et al. A Strategy for Automatic Moderation of a Large Data Set of Users Comments. In: *Computing Conference (CLEI), 2014 XL Latin American*. Montevideo, Uruguay: IEEE, 2014. p. 1–7.

SAYÃO, L. F.; SALES, L. F. Curadoria Digital: Um Novo Patamar para a Preservação de Dados Digitais de Pesquisa. *Informação & Sociedade*, v. 22, n. 3, p. 179–191, 2012.