Indexando Impressões Digitais Utilizando Índice Invertido: Uma Investigação Inicial

Johnny Marcos S. Soares[1], Luciano Barbosa[2], Paulo Antonio Leal Rego[3], Regis Pires Magalhães[1], Jose Antônio F. de Macêdo[3]

[1]Universidade Federal do Ceará - Campus Quixadá
Quixadá – CE – Brasil

[2]Centro de Informática – CIn
Universidade Federal de Pernambuco – Recife – PE – Brasil

[3]Departamento de Computação
Universidade Federal do Ceará – Fortaleza – CE – Brasil

johnnymarcos@alu.ufc.br,luciano@cin.ufpe.br,

pauloalr@ufc.br,{regis,jose.macedo}@insightlab.ufc.br

Abstract. *Fingerprints have been widely used for person identification. With the increase of fingerprint databases, indexing techniques are essential to perform efficient search over such great volume of data. This work is an initial attempt of leveraging well-established text indexing approaches and tools to fingerprint search. For that, our solution first converts fingerprints into text documents using techniques such as Minutia Cylinder-Code (MCC) and Locality-Sensitive Hashing (LSH), and then indexes them in inverted files using Elasticsearch, a highly scalable and distributed search engine.*

Resumo. *As impressões digitais têm sido amplamente usadas para identificação de pessoas. Com o aumento dos bancos de dados de impressões digitais, as técnicas de indexação são essenciais para realizar uma pesquisa eficiente em um volume tão grande de dados. Este trabalho é uma tentativa inicial de alavancar abordagens e ferramentas de indexação de texto bem estabelecidas para a pesquisa de impressões digitais. Para isso, nossa solução primeiro converte as impressões digitais em documentos de texto utilizando técnicas como Minutia Cylinder-Code (MCC) e Locality-Sensitive Hashing (LSH), e depois as indexa em arquivos invertidos usando o Elasticsearch, um mecanismo de pesquisa altamente escalável e distribuído.*

1. Introdução

As impressões digitais são uma das informações biométricas mais utilizadas em sistemas de verificação e identificação de pessoas, devido à imutabilidade, fácil aquisição e velocidade de processamento. Além disso, as impressões digitais possuem características distinguíveis entre indivíduos, sendo únicas até mesmo em gêmeos univitelinos [Maltoni et al. 2009]. A comparação de impressões digitais é realizada com o uso de algoritmos de *matching* [Maltoni et al. 2009]. Esses algoritmos utilizam informações extraídas das impressões digitais para compará-las e gerar um *score* de similaridade entre elas. Porém, a comparação de uma impressão digital de um individuo com uma grande

quantidade de outras impressões digitais pode tornar-se computacionalmente inviável, dado que o número de comparações pode chegar à ordem de milhares ou milhões. Com o objetivo de diminuir o espaço de busca de comparações de *matching*, são utilizadas técnicas de indexação de impressões digitais. Em alguns casos, as impressões digitais são agrupadas por informações como: dedo da impressão digital, qual das mãos possui aquele dedo, sexo da pessoa que possui a impressão digital, tipo de impressão digital, entre outras [Maltoni et al. 2009].

No entanto, devido ao grande aumento do número de impressões digitais tornou-se necessária a criação de abordagens mais robustas de indexação. As características, denominadas minúcias, são as informações mais utilizadas em técnicas de indexação de impressões digitais. De acordo com o padrão ANSI/NIST-ITL 2011, as minúcias possuem informações de posição, ângulo, qualidade e tipo [Mangold 2016]. O principal objetivo deste trabalho é realizar uma investigação inicial do uso de índices invertidos, amplamente utilizados para a indexação de texto, através da implementação de um método de indexação de impressões digitais baseado em minúcias. Para isso, utilizamos a técnica *Minutia Cylinder-Code* (MCC) [Cappelli et al. 2010b] que transforma uma minúcia em um vetor binário. Em seguida, é utilizado *Locality Sensitive Hashing* (LSH) [Datar et al. 2004] para gerar termos a partir de n-bits no vetor, criando assim um documento com esses termos. Os documentos gerados a partir desse processo são então indexados pelo Elasticsearch[1], um engenho de busca escalável para texto. Porém, pode ser utilizado outro engenho de busca textual.

O restante do artigo esta estruturado da seguinte forma: na Seção 2, o Elasticsearch é apresentado. Na Seção 3, é apresentada a proposta de indexação e busca. Na Seção 4, são mostrados os experimentos e resultados do método proposto. Na Seção 5, são apresentadas as considerações finais do trabalho.

2. Elasticsearch

A abordagem utilizada nesse artigo propõe o uso da escalabilidade e distribuição da ferramenta Elasticsearch para realizar a indexação e busca de impressões digitais. O Elasticsearch é um engenho de busca que utiliza índice invertido para a realização de indexação e busca, podendo chegar a *petabytes* de dados indexados [Gormley and Tong 2015]. Além disso, ele possui parâmetros para configurar um índice de maneira simples, como por exemplo, o parâmetro *shards* particiona o índice através dos nós no *cluster*. Outro parâmetro importante são as *réplicas* utilizadas no índice. Elas são cópias dos dados dos *shards* salvas em nós diferentes, garantindo mais disponibilidade aos dados. O Elasticsearch suporta consultas de documentos completos, denominada *full-text search*. Esse tipo de consulta utiliza modelos de similaridade para associar os documentos indexados com o documento buscado. O Elasticsearch possui disponíveis vários modelos tradicionais de Recuperação de Informação [Baeza-Yates et al. 1999] como por exemplo BM25 e similaridade de cosseno.

3. Sistema de Indexação Textual de Impressões Digitais

A Figura 1 apresenta a arquitetura do sistema implementado. Ele possui 3 componentes:

[1] *www.elastic.co*

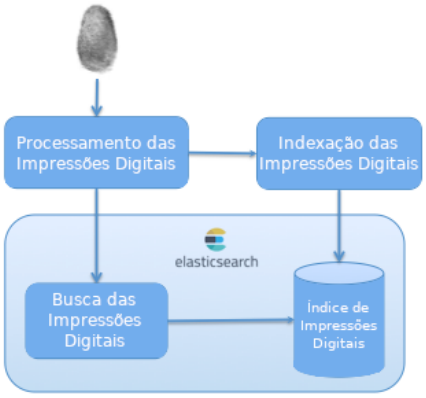


Figura 1. Arquitetura do sistema implementado

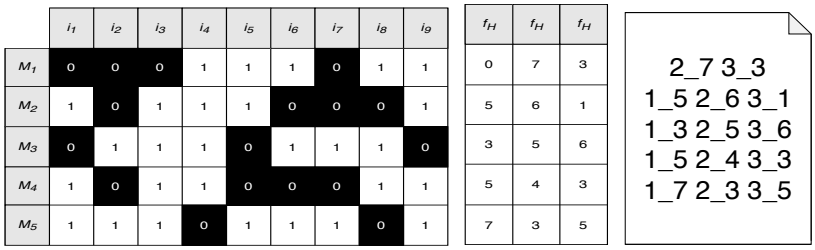
- Processamento das impressões digitais, responsável por processar as imagens das impressões digitais e transformá-las em documentos;
- Indexação de impressão digitais, que processa os documentos gerados pela etapa anterior e os indexa no Elasticsearch;
- Busca de impressões digitais é o componente dentro de Elasticsearch que realiza buscas por impressões digitais no índice.

Processamento das impressões digitais. O módulo de Processamento das Impressões Digitais realiza as seguintes etapas para transformar uma imagem de impressão digital em documento:

1. Dada a imagem da impressão digital, é utilizada a ferramenta MINDTCT do pacote NBIS (NIST Biometric Image Software) [Ko 2007] para extrair as minúcias das impressões digitais e suas informações de posição, ângulo e qualidade.
2. Em seguida, é utilizada a técnica Minutia Cylinder-Code (MCC) para gerar os vetores binários a partir do arquivo de minúcia. O MCC cria um vetor binário de tamanho n bits para cada minúcia encontrada.
3. Com os vetores binários criados, é aplicada em cada vetor uma função f_H que recebe um subvetor de h bits e retorna o valor decimal referente aos h bits. Em seguida, são criados e salvos em um documento, os termos referentes a cada termo gerado pela função f_H no formato k_b. No qual k é a posição do subvetor de h bits no vetor binário e b é o valor decimal retornado pela função f_H. Os termos que possuem o retorno da função f_H iguais a zero não são adicionados ao documento, devido à grande quantidade de sequências de zeros nos vetores binários.

Na Figura 2 é apresentado um exemplo do uso dessa estratégia. Na Figura 2(a) existem 5 vetores binários M_1, ..., M_5 de tamanho 9 bits. A Figura 2(b) mostra a utilização da função f_H, considerando o tamanho do subvetor h de 3 bits. Já na Figura 2(c) são mostrados os termos gerados a partir do retorno da função f_H, considerando a remoção dos termos com valor b=0. Outra característica importante utilizada nesta abordagem é a remoção de minúcias que possuem qualidade inferior a um limiar, pois essas podem degradar a qualidade do resultado das consultas. Além disso, o uso desse limiar diminui a quantidade de vetores binários gerados, reduzindo o tamanho do documento de termos e do tempo de processamento.

Indexação de Impressões Digitais. O Algoritmo 1 mostra os detalhes para a indexação de um conjunto de arquivos de minúcias já computadas de um conjunto de impressões digitais. O Elasticsearch pode realizar a indexação de documentos tanto individualmente



(a) Vetores binários do *Minutia Cylinder-Code* (b) Resultado da função f_H (c) Documento de termos gerados

Figura 2. Criação do documento de termos a partir dos vetores binários

quanto em conjuntos. A indexação individual é feita com a solicitação de inserção no Elasticsearch. Já a indexação em conjunto é feita com a utilização da *API Bulk*, no qual pode realizar tanto operações de inserção como de atualização e remoção de documentos em grandes quantidades [Gormley and Tong 2015].

Algoritmo 1: Indexação no Elasticsearch

Input: Um conjunto de arquivos com minúcias, $DB = \{F_1, F_2, ...\}$
Result: Indexação das informações textuais em um índice do Elasticsearch
for *cada arquivo de minúcia* F_i *em* DB **do**
  Use o MCC para criar os vetores binários V_i do arquivo de minúcia F_i
  for *cada vetor* v_j *em* V_i **do**
    for *cada subvetor* p_k *de* h *bits no vetor* v_j **do**
      $b = f_H(p_k)$
      if $b > 0$ **then**
        Insere o termo k_b no documento de termos referente ao arquivo V_i;
      end
    end
  end
  Insere o documento de termos com o identificador i no índice do Elasticsearch
end

Busca de Impressões Digitais. Dada a imagem de uma impressão digital que se deseja realizar uma busca no índice, nosso sistema inicialmente a processa utilizando o módulo de Processamento das Impressões Digitais, transformando-a em um documento texto. Em seguida, a consulta do documento é realizada utilizando a funcionalidade *search full-text*, mais especificamente, usando o operador *More Like This* (MLT). O MLT utiliza uma abordagem para limitar a quantidade de termos utilizados na consulta, reduzindo o tempo de processamento da busca. Por exemplo, a consulta com MLT automaticamente seleciona os termos com maior *tfidf* para realizar a busca. Após isso, são retornados os identificadores dos documentos mais semelhantes.

4. Experimentos e Resultados

Foram realizados experimentos com um conjunto privado de dados contendo 11 mil impressões digitais, na qual 10 mil são utilizadas na indexação e 1 mil na busca. As impressões digitais deste conjunto de dados são agrupadas a partir do valor da qualidade da impressão digital, que pode ser de 0 a 100 utilizando a ferramenta NFIQ 2 [Bausinger and Tabassi 2011]. Para esses experimentos, discretizamos esses valores em quatro intervalos: 0-25, 26-50, 51-75 e 76-100. A avaliação dos métodos de indexação de impressões digitais geralmente são realizadas com a relação entre *Penetration Rate* e *Error Rate*. *Penetration Rate* é a porcentagem da base que precisa ser pesquisada para encontrar o resultado correto. Já o *Error Rate* é a porcentagem de buscas que não obtiveram o resultado correto dentro do limite do *Penetration Rate*.

A Figura 3 mostra o gráfico de *Penetration Rate X Error Rate* do método proposto no conjunto de dados privado, com curvas referentes a grupos de qualidade das impressões digitais da busca. É possível perceber que quanto maior a qualidade da imagem melhor é o desempenho da busca. A curva referente às impressões digitais com maior qualidade (valores de qualidade entre 76 e 100) possui *Error Rate* de 0%, ou seja, todas as impressões digitais buscadas foram encontradas. Como nosso objetivo é lidar com escalabilidade e, consequentemente, prover buscas eficientes, um outro fator importante na análise da solução é o tempo de resposta. O tempo médio por consulta nesses experimentos foi de 0,2 segundos, o que confirma a eficiência de nossa abordagem para busca. Entretanto, não foi possível verificar o tempo de busca de outros métodos no conjunto de dados utilizado, pois os códigos-fonte não foram publicados.

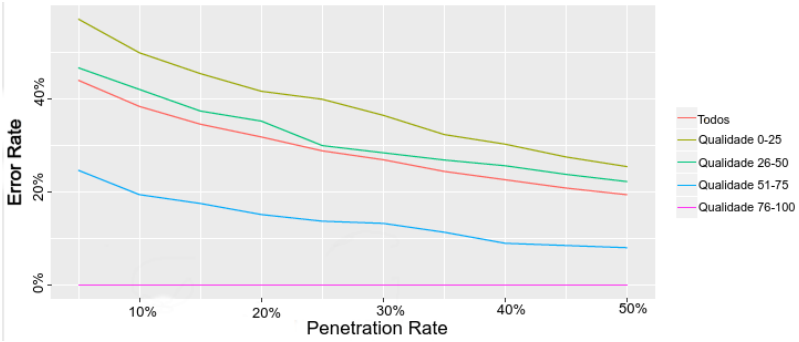


Figura 3. Gráfico de desempenho do método proposto

Outros dois conjuntos de dados públicos e menores foram utilizados (FVC2000 DB2a [Maio et al. 2002a] e FVC2002 DB1a [Maio et al. 2002b]) para comparação com a técnica de indexação de impressões digitais proposta por [Cappelli et al. 2010a], dado que ela utilizou essas bases para avaliação. Ambas possuem 800 imagens de impressões digitais de 100 dedos, totalizando 8 imagens por dedo. Uma impressão digital de cada dedo é considerada principal e será indexada no Elasticsearch, já as outras 7 são utilizadas para realizar as buscas. Com a FVC2000 DB2a, o método proposto obteve 22,8% de *Error Rate* com 10% de *Penetration Rate* e 8,7% de *Error Rate* com 10% de *Penetration Rate* na FVC2002 DB1a. Já o método proposto em [Cappelli et al. 2010a] obteve 2% de *Error Rate* e 1% *Error Rate* com 10% de *Penetration Rate* nas bases FVC2000 DB2a e FVC2002 DB1a, respectivamente. Na Tabela 1 são apresentados os parâmetros que obtiveram o melhor resultado nos experimentos do método proposto neste trabalho utilizando dados públicos. O método proposto obteve, portanto, um resultado inferior ao trabalho comparado, porém, os métodos de indexação de impressões digitais que utilizam minúcias realizam buscas individuais por cada minúcia no conjunto de dados, já o método proposto realiza uma única busca utilizando todas as minúcias. Essa generalização das minúcias diminui a qualidade do resultado, porém, torna-se mais rápida em ambientes com uma grande quantidade de dados.

Parâmetro	Descrição	Valor em FVC2002 DB1a	Valor em FVC2000 DB2a
t	Quantidade de termos usados no MLT	250	100
n	Tamanho do vetor binário	1944	1944
h	Quantidade de bits usados em f_H	8	12
T_q	Limiar de qualidade das minúcias	15	35

Tabela 1. Valores dos melhores parâmetros usados nos experimentos

5. Conclusão

Foi proposto neste trabalho um novo método de indexação de impressões digitais baseado em *Minutia Cylinder-Code* e na utilização do Elasticsearch. Essa abordagem utiliza métodos empregados em buscas textuais no domínio de impressões digitais. Portanto, com o uso do método proposto é possível indexar uma grande quantidade de documentos gerados a partir de impressões digitais e aproveitar a fácil escalabilidade e distribuição do Elasticsearch. Como trabalho futuro, a indexação de documentos no Elasticsearch pode ser realizada à nível de minúcia, ao invés de indexar um documento por impressão digital. Essa modificação pode melhorar a qualidade do resultado, pois grande parte das técnicas baseadas em minúcias utilizam essa abordagem.

Agradecimentos. Os autores gostariam de agradecer à Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP) pelo apoio financeiro (Processo 8789771/2017).

Referências

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Bausinger, O. and Tabassi, E. (2011). Fingerprint sample quality metric nfiq 2.0. *BIOSIG 2011–Proceedings of the Biometrics Special Interest Group*.

Cappelli, R., Ferrara, M., and Maltoni, D. (2010a). Fingerprint indexing based on minutia cylinder-code. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):1051–1057.

Cappelli, R., Ferrara, M., and Maltoni, D. (2010b). Minutia cylinder-code: A new representation and matching technique for fingerprint recognition. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2128–2141.

Datar, M., Immorlica, N., Indyk, P., and Mirrokni, V. S. (2004). Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262.

Gormley, C. and Tong, Z. (2015). *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. O'Reilly Media, Inc.

Ko, K. (2007). User's guide to nist biometric image software (nbis). Technical report.

Maio, D., Maltoni, D., Cappelli, R., Wayman, J. L., and Jain, A. K. (2002a). Fvc2000: Fingerprint verification competition. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):402–412.

Maio, D., Maltoni, D., Cappelli, R., Wayman, J. L., and Jain, A. K. (2002b). Fvc2002: Second fingerprint verification competition. In *Object recognition supported by user interaction for service robots*, volume 3, pages 811–814. IEEE.

Maltoni, D., Maio, D., Jain, A. K., and Prabhakar, S. (2009). *Handbook of fingerprint recognition*. Springer Science & Business Media.

Mangold, K. C. (2016). Data format for the interchange of fingerprint, facial & other biometric information ansi/nist-itl 1-2011 nist special publication 500-290 edition 3. Technical report.