colloquium exactarum

ANÁLISE VISUAL DA EVOLUÇÃO DE COLEÇÕES DE DOCUMENTOS UTILIZANDO TAG CLOUD

VISUAL ANALYSIS OF DOCUMENT COLLECTIONS EVOLUTION USING TAG CLOUD

André Luis Dias Andreotti[1]; Danilo de Medeiros Eler[1]; Ives Renê Venturini Pola[1]; Almir Olivette Artero[1];

Faculdade de Ciências e Tecnologia – Universidade Estadual Paulista "Júlia de Mesquita Filho" (FCT/UNESP)[1]
dias.andreotti@gmail.com; daniloeler@fct.unesp.br; ivesrene@gmail.com; almir@fct.unesp.br;

RESUMO – Nos dias atuais, há uma grande quantidade de informações disponíveis em diversas coleções de documentos, dentre eles artigos, teses, e outros trabalhos acadêmicos. Coleções de artigos científicos podem representar a trajetória acadêmica de um pesquisador em particular, e a evolução de uma linha de pesquisa com o passar dos anos. Neste trabalho, são investigadas técnicas de exploração e de visualização de documentos, com foco na técnica tag cloud. Para tanto, foi desenvolvida uma ferramenta que auxilia na visualização de coleções de documentos, possibilitando verificar a evolução da coleção ao longo do tempo, quanto o surgimento de novos tópicos ou quanto ao seu conteúdo, comparando a nuvem de etiquetas que foi gerada para cada documento.
Palavras-chave: tag cloud; documentos; evolução.

ABSTRACT - Nowadays, there is a lot of information available in various document collections, including articles, theses, and another scholarly works. Collections of papers may represent the academic career of a researcher in particular, and the development of a research line over the years. In this work, we investigated techniques of exploration and visualizations of documents, focusing on tag cloud technique. Therefore, a tool was developed to aid in document collection visualization and exploration, enabling to check the collection evolution over time, regarding to the emergence of new topics or their content, comparing distinct tag clouds generated for each document.
Keywords: tag cloud; documents; evolution.

1. INTRODUÇÃO

Atualmente, existe um amplo volume de dados armazenados por diversos meios eletrônicos. A exploração e análise desse vasto volume de dados está se tornando cada vez mais difícil, devido à grande quantidade de dados e de atributos. Por isso, muita pesquisa tem sido realizada em técnicas que apoiem os usuários no entendimento desses dados. Uma das áreas de pesquisa que tem se destacado nos últimos anos é a análise visual de dados. Na literatura, há um grande número de técnicas de visualização de informação que foram desenvolvidas ao longo da última década para apoiar a exploração de grandes conjuntos de dados. Entretanto, a habilidade do ser humano de armazenar dados aumenta a uma velocidade mais rápida do que a capacidade de analisá-los e, se nenhuma informação for extraída, o armazenamento terá sido inútil, afirma Keim (2002).

Um exemplo de conjunto de dados que tem crescido nos últimos anos são as coleções de artigos científicos, as quais podem representar a trajetória acadêmica de um pesquisador em particular, e a evolução de uma linha de pesquisa com o passar dos anos. Assim, analisar a evolução temporal dos assuntos abordados em coleções como estas é de extrema importância para acadêmicos e pesquisadores em geral, pois, podem avaliar as tendências da área que pretendem estudar e tomarem decisões a partir desta avaliação. Desta forma, é notável o quão útil é propor ferramentas que auxiliem nesta análise, facilitando a exploração textual.

Este trabalho trata essencialmente do estudo e aplicação da técnica de Visualização de Informação *tag cloud*, visando a exploração textual de uma coleção de documentos na busca pela evolução temporal de seus tópicos, tais como: identificação de emergências de tópicos em instantes específicos de tempo; a especialização de um tópico; e o desaparecimento de tópicos. Essa evolução temporal pode ser detectada por meio da identificação de termos mais relevantes de cada documento, analisando as nuvens de etiquetas geradas.

Neste trabalho foi desenvolvida a ferramenta TGEE (*Tag Clouds Enviroment Exploration*), com objetivo de facilitar a análise desse amplo volume de dados textuais.

O restante deste trabalho está organizado em cinco seções, na Seção 2 é realizada uma breve contextualização da área estudada para o desenvolvimento do presente trabalho. Na Seção 3 é descrita a abordagem adotada para o desenvolvimento da ferramenta TGEE (*Tag Clouds Enviroment Exploration*). Na Seção 4 são descritos os

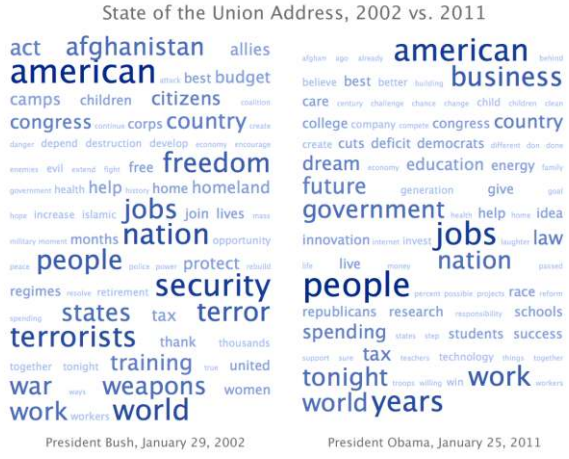
experimentos realizados e os resultados obtidos são exibidos. Finalmente, na Seção 5 as considerações finais e conclusões são apresentadas, bem como sugestões de aperfeiçoamento e trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

A Visualização de Informação é uma área da Ciência da Computação que tem por objetivo dar sentido aos dados por meio de representações gráfica, com o intuito de ampliar a capacidade cognitiva do ser humano em processos de exploração de dados, conforme afirmam Card, Mackinlay e Shneidermann (1999), Chen (2006) e Spence (2007). Uma das técnicas da Visualização de Informação é a *tag cloud* (nuvem de etiquetas), a qual consiste na criação de uma nuvem de palavras, que pertençam a um documento, fazendo a distinção entre elas pelo tamanho da fonte ou cor. Assim, na representação gráfica gerada, as palavras são destacadas levando em consideração a quantidade de vezes que aparecem no texto, por intermédio da tokenização e contabilização de cada termo. Na Figura 1 são apresentadas duas visualizações geradas por meio da técnica *tag cloud*. As duas visualizações foram geradas pelo site *tag crowd*, cuja *URL* é www.tagcrowd.com. A visualização da esquerda foi processada utilizando o texto do discurso realizado pelo presidente dos Estados Unidos da América, George W. Bush, no ano de 2002. E a outra visualização sobre o discurso realizado pelo, até então, atual presidente Barack Obama em 2011.

Palavras que foram mais pronunciadas pelos respectivos presidentes, foram destacadas nesta representação pelo tamanho da fonte e intensidade da cor azul. A partir da imagem é possível observar os assuntos que eram prioridade na época, em 2002, logo após o atentado terrorista de 11 de setembro de 2001, onde destacam as palavras "*american*", "*terrorists*", "*security*" e "*afghanistan*". Por meio dessa representação gráfica, é possível observar a troca de contexto nos assuntos políticos, quase dez anos após o discurso anterior.

Figura 1. Tag cloud comparando o discurso realizado pelos dois presidentes dos Estados Unidos da América, no ano de seus respectivos mandatos.



Fonte: TAGCROWD (2011).

A partir da visualização é possível ter uma ideia geral do assunto tratado nos dois discursos e as principais diferenças entre eles, sem que seja necessário ler todo o discurso. Apesar de parecer simples a ideia de que documentos distintos possuam distribuições de termos diferentes, esta hipótese mostrou-se bem-sucedida na classificação de textos em assuntos, mostrando desempenho superior ao de representações mais complexas, reforça Sebastiani (2002).

A Figura 2 mostra outro exemplo de visualização que utiliza tag cloud. Nesta representação a mudança positiva do preço da ação é indicada pela cor verde, a mudança negativa pela cor vermelha, e o tamanho da fonte indica a variação percentual.

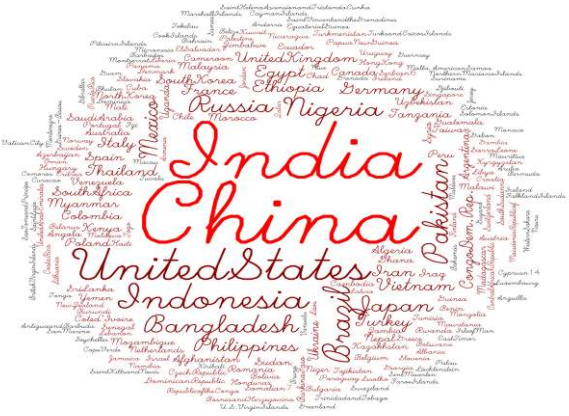
Figura 2. Movimento dos preços das ações.



Fonte: Warrena (2007).

Um exemplo que utiliza tag cloud para representar dados quantitativos está expresso na Figura 3, onde a população mundial foi distinguida pelo tamanho da fonte do nome do país.

Figura 3. População de cada um dos países do mundo. Criada em R (linguagem de programação) utilizando *wordcloud*.



Fonte: http://www.wordclouds.com

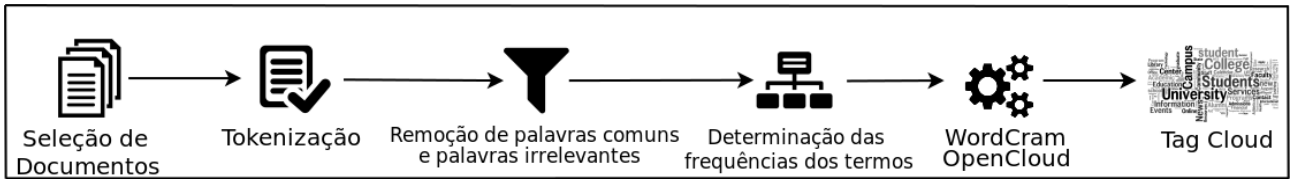
3. ABORDAGEM PROPOSTA

O processo de desenvolvimento da ferramenta TGEE (*Tag Clouds Enviroment Exploration*), fruto do presente trabalho, teve sua base nas etapas de pré-processamento e representação de coleções de documentos utilizadas por Alencar (2013), sendo elas: tokenização, remoção de stopwords, Stemming, Lei de Zipf e Corte de Luhn. A ferramenta foi desenvolvida utilizando a linguagem de programação *Java*, visando a garantia de portabilidade para sistemas operacionais que suportam a máquina virtual *Java* (*Java Virtual Machine – JVM*).

A Figura 4 descreve a abordagem proposta, mostrando cada etapa do processamento realizado, partindo da aquisição da coleção de documentos, passando pela tokenização dos termos, remoção de palavras comuns e irrelevantes

para a coleção analisada, contagem dos termos no modelo vetorial, utilização das bibliotecas *WordCram* e *OpenCloud* para, finalmente, gerar a visualização da *tag cloud*.

Figura 4. Etapas do processo da abordagem proposta.



Fonte: (Própria).

Foram utilizadas duas bibliotecas responsáveis por gerar a visualização das tag clouds, sendo elas: *WordCram* e *OpenCloud*. *WordCram* é um trabalho, disponível na *URL* wordcram.org, que utiliza a linguagem de programação *Processing* de código aberto e ambiente de desenvolvimento integrado, construído para as artes eletrônicas e comunidades de projetos visuais. *OpenCloud* **auxilia na projeção dos termos** horizontalmente no plano e em ordem alfabética, sendo menos sofisticada que a primeira, disponível na URL www.sourceforge.net/projects/opencloud/.

3.1 SELEÇÃO DE DOCUMENTOS

Visando a praticidade, no processamento de coleções de documentos pelo usuário final da ferramenta desenvolvida, foi implementada a possibilidade de seleção de vários arquivos textos, no formato "*txt*" (formato de arquivo somente texto, sem formatação), sendo possível optar por gerar uma *tag cloud* para cada arquivo selecionado, ou concatenar todo o texto dos arquivos e gerar apenas uma *tag cloud*.

3.2 TOKENIZAÇÃO

A primeira etapa a ser realizada é a tokenização do texto, separando-o em tokens. Caso o texto esteja devidamente estruturado, de acordo com a pontuação da linguagem utilizada, esta tarefa torna-se fácil, selecionando as palavras quando espaços ou caracteres de pontuação são encontrados. A Figura 5 exibe um exemplo deste processo.

Figura 5. Exemplo de tokenização de uma frase.

Entrada: "Você não pode juntar os pontos olhando para frente; você pode conectá-los apenas olhando para trás".
Saída: Você | não | pode | juntar | os | pontos | olhando | para | frente | você | pode | conectá-los | apenas | olhando | para | trás

Fonte: (Própria).

3.3 STOPWORDS

Uma etapa fundamental é a de verificar quais palavras representam melhor o assunto de um texto, normalmente, substantivos.

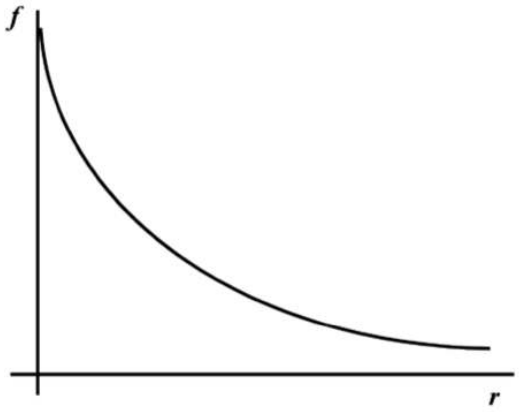
Assim, é necessário remover os termos irrelevantes, tais como artigos, adjetivos, advérbios, conjunções e preposições. Esses termos são conhecidos como *stopwords* e são dependentes da linguagem adotada no documento. Esse processo é feito por meio de listas pré-estabelecidas de *stopwords* da linguagem do documento processado.

3.4 Lei de Zipf e Cortes de Luhn

Deve-se considerar que termos que aparecem com muita frequência em vários documentos e outros que aparecem em poucos documentos podem ser desconsiderados no processamento, pois, não ajudam a diferenciar a semântica de uma coleção de documentos.

A Lei de Zipf (1949) leva em consideração o número de incidências de termos, e traça uma razão em comum entre eles. Considerando o histograma de frequência dos termos de uma coleção de documentos ordenado em ordem crescente obtém-se a "Curva de Zipf" observada na Figura 6. O eixo cartesiano "r" representa os termos, em ordem decrescente de frequência, e o eixo cartesiano "f" representa a frequência desses termos.

Figura 6. Curva de Zipf.

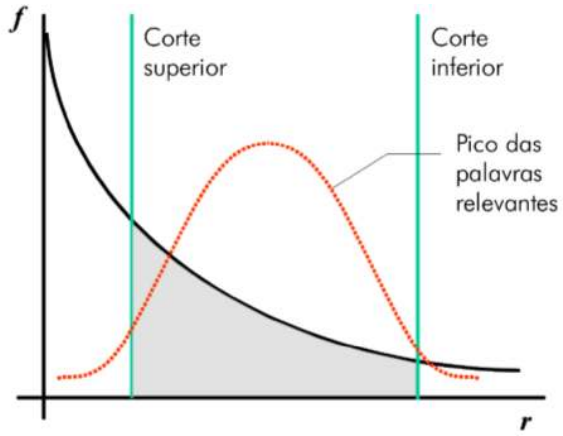


Fonte: Zipf (1949).

Luhn (1958), usou a Lei de Zipf para especificar dois pontos de corte, os quais denominou de superior e inferior, para separar os termos relevantes, observados na Figura 7. Desta forma, os termos que aparecem com muita frequência em vários documentos e outros que aparecem em poucos documentos, são desconsiderados. Com base neste conhecimento, a ferramenta TGEE considera apenas os termos mais significativos quando um documento é processado, diminuindo o esforço computacional necessário. É possível adicionar termos a uma lista de remoção, ignorando os que possuem uma alta frequência já esperada na coleção que será avaliada, por exemplo, ao processar uma coleção de documentos da área de Visualização de Informação, os termos: "informação" e "visualização" tendem a aparecer com uma alta frequência em todos os textos desta coleção, logo esses termos podem ser adicionados à lista de remoção,

pois, os mesmos não servirão para analisar a evolução temporal entre os documentos.

Figura 7. Cortes de Luhn.



Fonte: Luhn (1958).

3.6 TAG CLOUDS

O modelo vetorial, descrito em Salton, Wong e Yang (1975), também conhecido como *bag-of-words*, descreve cada documento como um vetor de frequências dos termos que nele ocorrem. Sendo $D = \{d_1, d_2, ..., d_N\}$ uma coleção de N documentos que inclui M termos $T = \{t_1, t_2, ..., t_M\}$. Cada documento d_i é um vetor v $(d_i) = \{freq_{i1}, freq_{i2}, ..., freq_{iM}\}$, no qual o valor $freq_{ij}$ é alguma medida que determina a influência do termo t_j no documento d_i, conforme representado na Tabela 1.

Tabela 1. Representação Vetorial.

	t_1	t_2	$\cdots$	t_M
d_1	$freq_{11}$	$freq_{12}$	$\cdots$	$freq_{1M}$
d_2	$freq_{21}$	$freq_{22}$	$\cdots$	$freq_{2M}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
d_N	$freq_{N1}$	$freq_{N2}$	$\cdots$	$freq_{NM}$

Fonte: Alencar (2013).

A frequência $freq_{ij}$ é calculada utilizando a medida *terms frequency* (tf). A medida considera o valor de ocorrências de t_j no documento d_i, sendo $freq_{ij}$ = tf (t_j, d_i) = freq (t_j, d_i).

No presente trabalho, para o armazenamento dos termos e respectivas frequências, de cada documento, foi utilizada a estrutura de dados árvore *rubro-negra*, contando com sua eficiência na realização de buscas e ordenação de elementos com uma complexidade assintótica de O (log (n)). Nesta fase do processamento, foi escolhida esta estrutura, pois é a etapa do pré-processamento que exige maior esforço computacional, visto que a cada termo retirado do texto deve-se verificar a lista de termos já extraídos e então caso já exista esse termo incrementa-se sua frequência, e caso seja um termo ainda não conhecido, adiciona-o na lista.

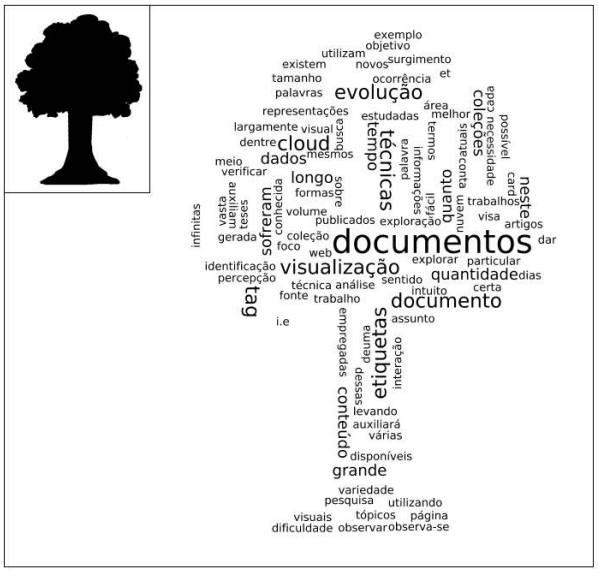
3.5 DETERMINAÇÃO DAS FREQUÊNCIAS DOS TERMOS

O objetivo principal da ferramenta TGEE é facilitar o entendimento de um conjunto de documentos por meio da técnica de

visualização tag cloud, exibindo a frequência dos termos mais relevantes de cada documento desta coleção. A partir da visualização é possível analisar quais termos foram surgindo com o passar dos anos e então definir a evolução temporal sofrida pelos documentos.

É possível alternar facilmente pelas duas técnicas, *WordCram* e *OpenCloud*, implementadas na ferramenta. Utilizando a visualização *WordCram* é possível carregar uma imagem que dará forma a disposição dos termos no plano, sendo que a técnica verifica os limites desta imagem carregada, e projeta os termos no mesmo formato. Aconselha-se utilizar uma imagem preenchida de preto e com fundo branco, para facilitar a identificação dos limites da mesma. Essa técnica exibe os termos distinguindo-os pelo tamanho da fonte e diferença de cores, logo a ferramenta desenvolvida permite a definição de cores e limites de tamanhos máximos e mínimos que serão utilizados nas fontes dos termos. Na Figura 8 observa-se um modelo de imagem de entrada, no canto superior esquerdo, e, em uma escala maior, é exibida a saída de um processamento realizado utilizando o texto de resumo deste artigo. Também é possível utilizar a técnica *OpenCloud*, nesta os termos são dispostos em ordem crescente e suas frequências representadas pelo tamanho da fonte, também é possível definir uma cor para os termos e uma cor de fundo. Nesta visualização não é necessário carregar uma imagem de formato, o custo computacional é bem menor comparado ao da *WordCram*, pois não há necessidade de verificar limites de imagem, e nem tratar sobreposições dos termos.

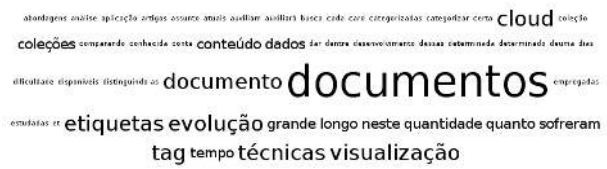
Figura 8. Imagem de entrada e saída do processamento, utilizando a técnica *WordCram*, sobre o texto de resumo deste próprio artigo.



Fonte: (Própria).

A Figura 9 mostra o resultado do processamento utilizando novamente o texto do resumo deste artigo. Na ferramenta é possível definir o diretório de saída onde serão armazenados os resultados dos processamentos, visualizar o vetor de frequência de cada documento processado e as estatísticas do processamento realizado.

Figura 9. Imagem de entrada e saída do processamento, utilizando a técnica *OpenCloud*, sobre o texto de resumo deste próprio artigo.



Fonte: (Própria).

4. EXPERIMENTOS

Para a validação da ferramenta TGEE descrita na seção anterior, foi analisada uma coleção de 96 trabalhos publicados pelos docentes do curso de Ciência da Computação da Faculdade de Ciência e Tecnologia (Unesp - Campus de Presidente Prudente, São Paulo, Brasil), nos últimos dez anos, de 2006 até 2016. As Figuras de 10 até 20 indicam o período analisado.

No ano de 2006 existe a publicação com o título: "*Empowering ISO-surfaces with volume data*" que trata especificamente de um algoritmo de visualização volumétrica denominado *VoS - Volume on Surface*. Observa-se na Figura 10 do respectivo ano, as palavras: "*volume*"; "*surface*"; e "*rendering*" em destaque. Mesmo sendo um único artigo, este utiliza uma alta frequência destes termos, desta forma os mesmos aparecem em destaque na visualização.

Percebe-se nas Figuras 11, 12 e 13, publicações dos anos 2007, 2008 e 2009, respectivamente, a alta frequência dos termos: "*bone*", "*age*", "*skeletal*" e "*hand*". Sendo que neste período existem vários trabalhos sobre métodos para estimativa da idade óssea, destacando-se a tese: "Novos métodos para estimativa da idade óssea baseado no processamento de imagens radiográficas da mão".

Existe na Figura 16, relacionada às publicações do ano de 2012, o destaque das palavras "*algorithm*" e "*panoramic*" sendo que neste ano existem duas publicações que tratam especificamente dos dois termos, sendo elas: "*Construction of Panoramic Images from Aerial Images Obtained by Aircrafts*" e "Um Novo Algoritmo para a Construção de Imagens Panorâmicas usando os algoritmos *Sift* e *Ransac*". Também, os termos "*mappings*", "*multiple*", e "*fibers*", relacionados a trabalhos de visualização.

Nos anos de 2013 e 2014, representados nas Figuras 17 e 18, respectivamente, observa-se em destaque o termo "*recognition*", e analisando os artigos publicados neste período são encontrados os seguintes títulos: "Detecção e Reconhecimento de Objetos em Imagens Utilizando Algoritmos de Extração De Pontos Chave"; "Reconhecimento de Objetos Coloridos e Mãos Usando Cores e Formas"; "Algoritmo para o Reconhecimento de Caracteres Manuscritos"; *"Traffic Sign Detection and Recognition using then AdaBoost and SIFT algorithm*"; e "Métodos

para reconhecimento de notas musicais de partituras: resultados preliminares". Sendo que todos estes artigos tratam de técnicas de reconhecimento.

A Figura 21 apresenta o resultado do processamento unindo todos os trabalhos publicados no período de 2006 a 2016, sendo possível observar que os trabalhos publicados neste período utilizam os temas: "*images*"; "*data*"; "*visualization*"; e "*recognition*". Na Figura 22 o mesmo processamento é realizado, porém são adicionadas na lista de remoção as palavras comuns da coleção processada, e outras que não são relevantes e aparecem em quase todos os trabalhos, por exemplo: "*work*"; "*approach*", e "*proposed*". Essa remoção foi realizada em todos os processamentos realizados, visando destacar palavras mais relevantes. Com a visualização é possível ter uma visão geral dos assuntos abordados nesta coleção de documentos, indicando as principais linhas de pesquisas dos professores do curso de Ciência da Computação, sendo elas Visão Computacional e Visualização. Também pôde-se observar o surgimento de alguns tópicos em épocas específicas. Por meio da ferramenta esta análise é feita sem a necessidade de uma leitura completa de cada publicação, mesmo considerando a utilização do resumo e palavras chaves para facilitar a leitura, ainda assim verifica-se o quão trabalhoso seria realizar esta tarefa.

Figura 10. *Tag Cloud* gerada a partir de seis publicações dos docentes do curso de Ciência Computação da FCT no ano de 2006.

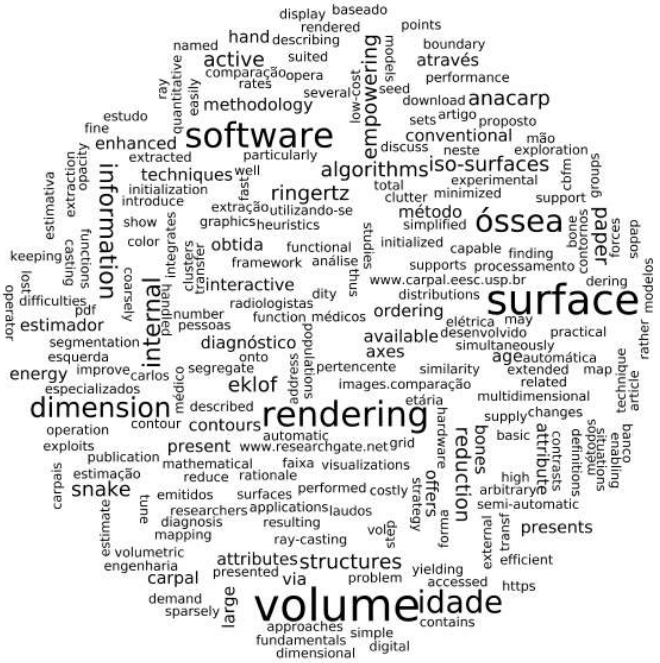
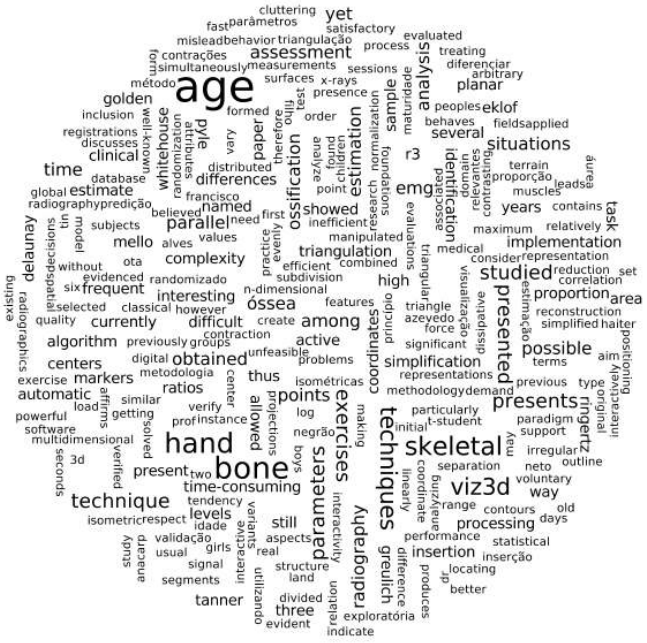

surface
rendering
volume
dimension
software
óssea
internal
information
estimador
diagnóstico
eklof
snake
ringertz
algorithms
iso-surfaces
anacarp
empowering
methodology
enhanced
techniques
interactive
available
ordering
reduction
bones
attributes
structures
presents
carpal
present
obtida
método
através
active
hand
display
baseado
rendered
points
boundary
performance
download
artigo
proposto
conventional
discuss
neste
exploration
experimental
minimized
support
simplified
initialized
capable
finding
processamento
distributions
www.carpal.eesc.usp.br
functional
framework
análise
radiologistas
médicos
pessoas
function
number
segmentation
esquerda
energy
especializados
operation
contour
contours
automatic
www.researchgate.net
grid
mathematical
faixa
visualizations
reduce
rationale
emitidos
surfaces
performed
costly
researchers
applications
laudos
diagnosis
resulting
mapping
ray-casting
volumetric
engenharia
presented
via
problem
yielding
accessed
https
demand
sparsely
large
approaches
fundamentals
dimensional
simple
digital
contains
idade
axes
age
automática
extended
map
similarity
pertencente
images.comparação
related
multidimensional
etária
changes
supply
basic
contrasts
definitions
arbitrarye
semi-automatic
high
offers
strategy
forma
transf
external
efficient
publication
estimação
carpais
exploits
estimate
tune
volts
graphics
heuristics
utilizando-se
extração
introduction
initialization
extracted
particularly
well
last
total
clutter
cbfm
groups
contornos
forces
sopep
dados
modelos
rather
technique
article
banco
labelling
segmentation
simultaneously
practical
desenvolvido
elétrica
may
dering
keeping
functions
casting
coarsely
handled
integrates
clusters
transfer
lost
operator
difficulties
pdf
estudo
fine
opacity
extraction
estimativa
show
color
obtida
suited
comparação
opera
rates
several
seed
low-cost
models
named
quantitative
ray
easily
sets
mão
groups
cbfm
experimental
minimized
support
forces
sopep
dados
modelos
rather
technique
article
banco
labelling
definitions
contrasts
basic
supply
changes
multidimensional
related
extended
map
similarity
pertencente
images.comparação
etária
hardware
reduction
bones
attributes
structures
presents
high
arbitrarye
semi-automatic
offers
strategy
forma
transf
external
efficient
yielding
accessed
https
problem
via
presented
carpal
volumetric
engenharia
estimate
tune
volts
ray-casting
mapping
diagnosis
resulting
researchers
applications
laudos
emitidos
surfaces
performed
costly
reduce
rationale
mathematical
faixa
visualizations
publication
estimação
carpais
exploits
operation
contour
contours
automatic
www.researchgate.net
grid
present
dimension
rendering
eklof
snake
energy
improve
carlos
médico
segregate
onto
esquerda
segmentation
estimador
diagnóstico
axes
age
automática
extended
map
similarity
pertencente
images.comparação
related
multidimensional
etária
changes
supply
basic
contrasts
definitions
arbitrarye
semi-automatic
high

Figura 11. *Tag Cloud* gerada a partir de cinco publicações dos docentes do curso de Ciência Computação da FCT no ano de 2007.



Fonte: (Própria).

Figura 12. *Tag Cloud* gerada a partir de três publicações dos docentes do curso de Ciência Computação da FCT no ano de 2008.



Fonte: (Própria).

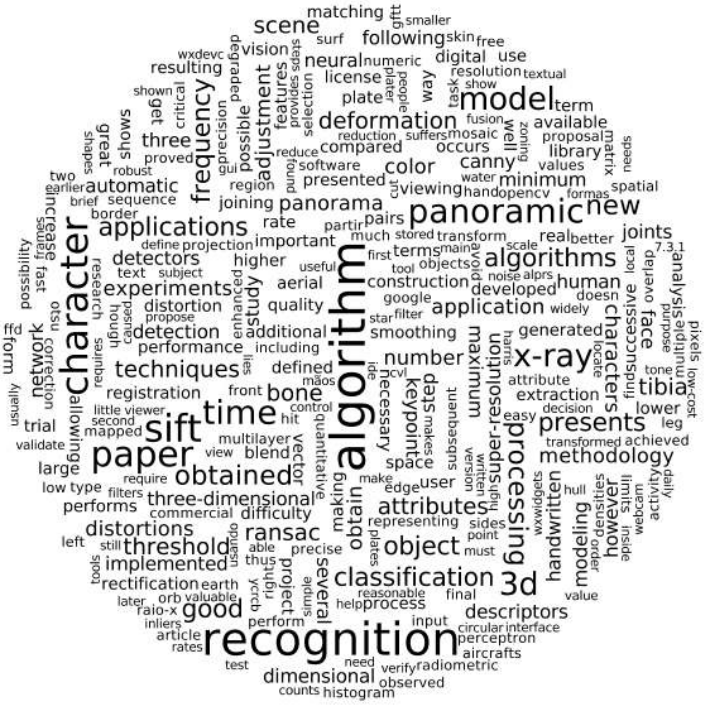
Figura 13. *Tag Cloud* gerada a partir de nove publicações dos docentes do curso de Ciência Computação da FCT no ano de 2009.


bone
analysis
age
eklof
hand
terrain
methodology
ringertz
representation
skeletal
collections
software
exploration
estimation
presents
detection
document
measurements
assessment
multidimensional
simplifications
application
paper
integration
classical
opensource
ossification
averages
digital
complexity
features
maturidade
among
techniques
mappings
growth
bones
centers
ages
present
time
task
test
idade
structured
radiograph
personal
children
multiple
producing
structures
projection
modeling
direction
configuration
radiologist
surface
topic
performed
coordination
strategy
programming
maturidade
exploration
reports
text
high
pediatric
usually
employed
standard
feature
problem
triangular
significant
timeconsuming
unstructured
crucial
lack
old
show
vertex
widely
aim
propose
thus
set
therefore
directions
known
complex
analyze
obtained
radiographic
various
sixteen
available
one
linear
three
aiming
makes
topics
qt
layout
planar
non-maximum
nine
left
tissue
cgal
ogr
mode
problems
relieffs
since
six
previously
regression
traditional
files
metacarpals
introduces
humans
opengl
surfaces
ratio
plane
follows
products
demo
variable
reduction
issues
two
t-student
projection
structures
varying
freely
medical
showing
synthetic
vision
maximal
costs
good
pediatric
known
complex
directions
high

Figura 15. *Tag Cloud* gerada a partir de oito publicações dos docentes do curso de Ciência Computação da FCT no ano de 2011.


projection
paper
spatial
applications
support
features
analysis
techniques
characters
presents
special
morphing
aspect-oriented
technique
interactive
parameters
distribution
representation
focusing
allowing
properly
models
continued
physical
aspects
dtms
space
gráfica
television
either
shape
enabling
outcome
user
classes
creation
tasks
software
artificial
plant
represents
surface
range
places
statistical
impossibilities
piecewise
mainly
variable
makes
fact
wide
process
useful
terrestrial
three
rely
verify
way
task
visualization
versatileapproach
effect
build
test
relevant
actors
scenes
laplacian-based
property
usefulness
made
character
major
still
neural
pixel
shows
supported
relative
application
estimates
supervised
movies
many
usually
therefore
thus
equal
geostatistical
understand
programs
visually
pixels
dados
simulation
guide
values
multidimensional
cause
selection
long
semivariograms
natural
especiais
pictures
class
levels
define
kriging
performs
flexible
generate
order
recognition
feature
impossible
increases
terms
continuity
variation
details
interpolation
according
information
model
graphics
krigagem
dataset
mechanism
set
unique
experiments
proposal
suitable
risks
essential
classification
precision
show
two
improving
important
terrenos
presented
feature
mapping
técnicas
sparsely
interaction
alter
novel
strategy
code
speed
well
local
large
physiology
plantas
purely
laws
real
understanding
propose
visualização
gráfica
species
space
dtms
terrain
representation
impossibilities
mainly
variable
makes
fact
piecewise
present
distribution
parameters
enabling
statistical
creation
tasks
software
outcome
usual
user
classes
plant
places
artificial
represents
surface
range
useful
wide
process
verify
way
task
visualization
versatileapproach
terrestrial
three
rely
form
map
special
either
shape
morphing
impossibilities
mainly
variable
makes
fact
enabling
statistical
creation
tasks
software
aspect-oriented
representation
parameters
distribution
piecewise
present
aspects
dtms
space
continued
physical
effect
build
test
relevant
actors
scenes
laplacian-based
property
usefulness
made
character
major
still
neural
pixel
shows
supported
relative
application
estimates
supervised
movies
many
usually
therefore
thus
equal
geostatistical
understand
programs
visually
pixels
dados
simulation
guide
values
multidimensional
cause
selection
long
semivariograms
natural
especiais
pictures
class
levels
define
kriging
performs
flexible
generate
order
recognition
feature
impossible
increases
terms
continuity
variation
details
interpolation
according
information
model
graphics
krigagem
dataset
mechanism
set
unique
experiments
proposal
suitable
risks
essential
classification
precision
show
two
improving
important
terrenos
presented
feature
mapping
técnicas
sparsely
interaction
alter
novel
strategy
code
speed
well
local
large
physiology
plantas
purely
laws
real
understanding
propose
visualização
gráfica
species
space
dtms
terrain
representation
impossibilities
mainly
variable
makes
fact
piecewise
present
distribution
parameters

Figura 17. *Tag Cloud* gerada a partir de 13 publicações dos docentes do curso de Ciência Computação da FCT no ano de 2013.



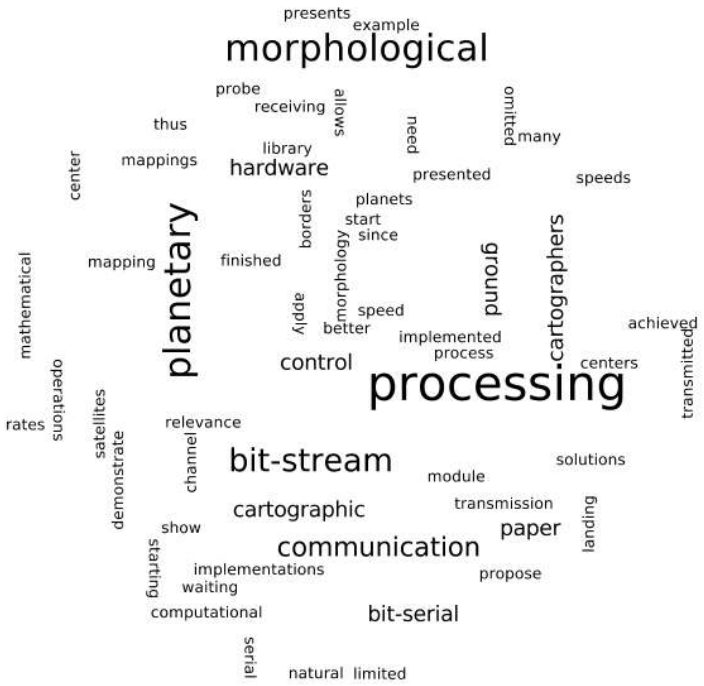
Fonte: (Própria).

Figura 18. *Tag Cloud* gerada a partir de 28 publicações dos docentes do curso de Ciência Computação da FCT no ano de 2014.



Fonte: (Própria.)

Figura 19. *Tag Cloud* gerada a partir de uma publicação dos docentes do curso de Ciência Computação da FCT no ano de 2015.



Fonte: (Própria).

Figura 20. *Tag Cloud* gerada a partir de duas publicações dos docentes do curso de Ciência Computação da FCT no ano de 2016.



Fonte: (Própria).

Figura 21. *Tag Cloud* gerada a partir de 96 publicações dos docentes do curso de Ciência Computação da FCT no período de 2006 a 2016.



Fonte: (Própria).

Figura 22. *Tag Cloud* gerada a partir de 96 publicações dos docentes do curso de Ciência Computação da FCT no período de 2006 a 2016. Removendo palavras comuns da coleção e palavras irrelevantes.



Fonte: (Própria).

5. CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho foi estudada e aplicada a técnica de Visualização de Informação *tag cloud* na busca pela exploração de coleções de documentos e análise da evolução temporal sofrida por seus termos. Para a avaliação das técnicas investigadas foi implementada uma ferramenta, denominada TGEE, visando facilitar a exploração de coleções de documentos.

Para validação da ferramenta foi realizado o processamento numa coleção de artigos publicados pelos professores do curso de Ciência da Computação da Faculdade de Ciência e Tecnologia (Unesp - campus de Presidente Prudente). Pôde-se alcançar os resultados esperados, sendo possível verificar os principais tópicos pesquisados pelos docentes da computação do campus. O usuário pode escolher e alternar facilmente entre as duas técnicas implementadas, sendo elas *WordCram* e *OpenCloud*. Ainda persistem na visualização algumas palavras irrelevantes, porém a visualização ainda destaca as mais relevantes.

As duas técnicas utilizadas possuem suas particularidades, sendo a *WordCram* mais robusta, possuindo um visual elegante, possibilitando uma customização pelo usuário, por meio da escolha de imagem de formato, em contrapartida possui um maior custo computacional para a execução de seu processamento. Já a *OpenCloud* conta com uma visualização limpa, para o usuário que almeje uma visualização mais simples e um processamento menos custoso, exibi os termos em ordem alfabética, horizontalmente no plano, um após o outro.

Por meio dos experimentos realizados foi possível identificar que alguns pontos podem ser aperfeiçoados, por exemplo, algumas palavras irrelevantes que ainda persistem após o pré-processamento de remoção de palavras comuns, sendo que estas não são importantes para a análise do texto, também existem algumas palavras que estão presentes no documento tanto no plural quanto no singular, por exemplo, "*image*" e "*images*" na Figura 21, que poderiam ser unidas e representadas de forma única na visualização, aumentando suas frequências, levando em consideração que são equivalentes quanto ao seu significado.

Cabe ressaltar, que alguns termos aparecem com uma alta frequência em anos específicos, entretanto, quando se gera a tag cloud de toda a coleção, estes termos aparecem pequenos, ou nem aparecem, na visualização. Isso ocorre, pois, a frequência destes termos são relevantes em comparação aos demais termos do mesmo ano, porém, comparando-os com a coleção inteira possuem uma baixa frequência. Entretanto, pode-se ser implementada uma

normalização, das quantidades por ano, destes termos, comparando a frequência máxima de cada ano e atribuindo um peso a cada termo, para que os mesmos tenham uma significância maior na tag cloud gerada para todos os anos.

Futuramente, a ferramenta poderá ser expandida para aceitar como entrada coleções de documentos em vários formatos que possuam larga utilização no meio acadêmico, por exemplo, *ISI*, *Endnote Export Format* ou *BibTeX*, facilitando a inserção dos dados. Também pode-se ser considerado a similaridade dos termos para unir termos que possuam o mesmo significado, e a utilização de uma matriz de covariância para detectar a frequência conjunta dos termos em todos os documentos.

REFERÊNCIAS

ALENCAR, A. B. Visualização da evolução temporal de coleções de artigos científicos. 2013. Tese (Doutorado) - Instituto de Ciências Matemáticas e de Computação (USP), São Carlos-SP, 2013.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMANN, B. Readings in information visualizations: using vision to think. San Francisco, CA: Morgan Kaufmann, 1999. 712 p.

CHEN, C. Information visualization: beyond the horizon. Secaucus, NJ: Springer-Verlag, 2006.

KEIM, D. A. Information visualization and visual data mining. IEEE Transactions in Visualization and Computer Graphics, v. 8, n. 1, p. 1-8, jan. 2002. https://doi.org/10.1109/2945.981847

LUHN, H. P. The automatic creation of literature abstracts. IBM Journal of Research and Development, Riverton, v. 2, n. 2, p. 159-165, abr. 1958.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. Communications of the ACM, New York, v. 18, n. 11, p. 613-620, nov. 1975. https://doi.org/10.1145/361219.361220

SEBASTIANI, F. Machine learning in automated text categorization. ACM Computing Surveys, v. 34, n. 1, p. 1-47, mar. 2002. https://doi.org/10.1145/505282.505283

SPENCE, R. Information visualization: design for interaction. 2. ed. Harlow, England: Prentice Hall, 2007.

TAGCROWD visualization: State of the Union. 2011. Disponível em: < http://tagcrowd.com/blog/2011/03/05/state-of-the-union-2002-vs-2011/>.

WARRENA. Data cloud graph showing the closing percentage increase or decrease of the New York Stock Exchange. 2007.

ZIPF, G. K. Human behaviour and the principle of least effort: an introduction to human ecology. Cambridge, MA: Addison-Wesley, 1949. 578 p.