Chronos - Um Motor de Busca Temporal

Aline Frozi[1], Edimar Manica[1,3], Renata Galante[1], Carina F. Dorneles[2]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil

[2]Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Campus Universitário Trindade – 88.049-900 – Florianópolis – SC – Brasil

[3]Instituto Federal do Rio Grande do Sul (IFRS) – Campus Avançado Ibirubá
Rua Nelsi Ribas Fritsch, nº 1111, Bairro Esperança – 98.200-000 – Ibirubá – RS – Brasil

{arfrozi,emanica,galante}@inf.ufrgs.br, dorneles@inf.ufsc.br

Resumo. *Este artigo apresenta a ferramenta Chronos, que oferece uma interface Web que possibilita aos usuários a realização de consultas temporais por palavras-chave em páginas Web. A principal contribuição da ferramenta é identificar os termos relacionados com cada expressão temporal presente no conteúdo das páginas Web. Experimentos iniciais mostram que Chronos melhora a qualidade dos resultados para consultas temporais em comparação com um motor de busca convencional.*

Abstract. *This paper presents a tool, named Chronos, that provides a Web interface that enables users to perform temporal queries for keywords on Web pages. The main contribution of the tool is to identify the terms related to each temporal expression contained in the Web pages content. Initial experiments show that Chronos improves the results quality for temporal queries when compared with a conventional search engine.*

1. Introdução

As páginas Web descrevem inúmeros tipos de informações, tais como política, economia e saúde. Muitas destas informações descrevem eventos que se repetem ao longo do tempo. Por exemplo, a Copa do Mundo ocorre a cada 4 anos. Muitas pessoas que pesquisam por este tema estão interessadas em informações sobre a copa atual. No entanto, existem pessoas interessadas em saber o desempenho de determinadas equipes em copas passadas ou informações sobre a copa de 2014 no Brasil.

Motores de busca comerciais têm explorado o valor da informação temporal na busca Web. Por exemplo, o Google permite definir o período de interesse, bem como a opção de visualizar os resultados da busca em uma linha de tempo. No entanto, estes motores de busca apenas consideram a data de coleta das páginas [Jin et al. 2008]. Entretanto, quando a página contém informações relacionadas ao passado ou ao futuro, há uma lacuna entre o tempo de coleta da página e o tempo a que a informação se refere. Há trabalhos que propõem motores de busca temporal. Porém, não são adequados para páginas que descrevem várias informações temporais (conforme é discutido na Seção 4).

Um problema que surge com a utilização de expressões temporais presentes no conteúdo das páginas Web é identificar quais termos da página se referem a cada expressão temporal. Por exemplo, consideram-se os trechos abaixo como dois exemplos

de páginas e a seguinte consulta "`Campeão Mundial Futebol 1930`", cujo objetivo é procurar o país vencedor da Copa do Mundo de 1930. Sem a identificação de quais termos se referem a cada expressão temporal, será atribuído o mesmo *score* de relevância para as duas páginas, pois ambas contêm as palavras-chave da consulta, inclusive a expressão temporal (`1930`). Porém, como pode-se observar, a `Página 1` não informa quem foi o Campeão Mundial em 1930.

Página 1. Em **1930**, o Brasil esteve presente para a disputa da 1ª Copa do Mundo de futebol... Em **1958**, o Brasil sagrava-se pela 1ª vez Campeão Mundial de Futebol...

Página 2. A Copa do Mundo de **1930** foi a primeira Copa do Mundo de Futebol realizada. Em **30 de julho de 1930**, o Uruguai se tornou o 1º Campeão Mundial de Futebol...

Este artigo descreve a ferramenta **Chronos**, que é um motor de busca temporal que permite a realização de consultas temporais por palavras-chave sobre a informação temporal presente no conteúdo das páginas Web. As principais contribuições da ferramenta são: (i) o suporte a um conjunto de consultas temporais; (ii) o tratamento da informação temporal presente no conteúdo das páginas Web, identificando quais termos se referem a cada expressão temporal – permitindo tratar páginas que contêm diversas informações temporais; e (iii) o método de indexação temporal.

O artigo está organizado como segue. Na Seção 2 é apresentada a ferramenta **Chronos**. A Seção 3 descreve os experimentos realizados e os resultados obtidos. Os trabalhos relacionados são apresentados na Seção 4. Finalmente, na Seção 5, são descritas as considerações finais e as direções futuras.

2. Ferramenta Chronos

A ferramenta **Chronos** é utilizada da seguinte forma. O usuário interage com uma interface fornecendo uma consulta temporal. Esta consulta temporal é constituída de palavras-chave convencionais e expressões temporais. Por exemplo, "`senator Brazil intersects 2005`" . Esta consulta é enviada a um motor de busca. O motor de busca, por meio de seu módulo de *ranking*, processa a consulta, acessa um índice invertido convencional e um índice temporal e retorna os resultados ranqueados de acordo com uma média ponderada entre a relevância do texto, a relevância do tempo e a importância da página. O motor de busca é uma adaptação do motor de busca Lucene (http://lucene.apache.org). A Subseção 2.1 descreve a especificação das consultas aceitas pela ferramenta. A Subseção 2.2 descreve o processo de indexação. A Subseção 2.3 descreve o módulo de *ranking*. A Subseção 2.4 apresenta a interface da ferramenta.

2.1. Especificação das Consultas

A forma genérica[1] de uma consulta temporal possibilitada pela ferramenta segue o formato definido em [Manica et al. 2010]:

```
tempQuery ::= {kws} tempPredicate
kws ::= string
tempPredicate ::= tempOp op
tempOp ::= ( After | Before | Contains | During | Equals | FinishedBy | Finishes |
             Intersect | Meets | MetBy | OverlappedBy | Overlaps | StartedBy | Starts )
op ::=  ( instant | interval | compInterval )
instant ::= [ number / ] [ number / ] number
interval ::= [ instant , [ instant ] ] | [ , instant ]
compInterval ::= word  [value] granularity
word ::= ( current | last | next )
granularity ::= ( Day(s) | Week(s) | Month(s) | Year(s) | Decade(s) | Century(ies) )
value ::= integer
```

[1] colchetes [] indicam segmento opcional, chaves {} indicam segmento repetitivo opcional, parênteses indicam um conjunto de opções, | indica "ou", e palavras em negrito são usadas como termos pré-definidos.

Uma consulta temporal (`tempQuery`) é composta por: um conjunto de palavras-chave convencional sem informação temporal (`kws`) e um predicado temporal (`tempPredicate`). O predicado temporal é uma restrição temporal dividida em duas partes: operador temporal (`tempOp`), e operando (`op`). O operando é um instante ou um intervalo, que deve seguir o operador temporal. O operador temporal define a relação temporal entre o conjunto de palavras-chave convencional e o operando. É importante notar que o operador temporal cobre todas as relações básicas entre intervalos temporais definidas em [Allen 1983], mais o operador INTERSECT.

Há três possibilidades para usar o operando: instante (`instant`), intervalo (`interval`) ou intervalo composto (`compInterval`). O instante é um ponto no tempo. Por exemplo, `2008`. O intervalo é definido por um período temporal composto por dois instantes que representam, respectivamente, o instante inicial e final do período. Estes dois instantes devem estar entre colchetes e separados por vírgula. Por exemplo, `[01/01/2008, 30/06/2008]`. Um intervalo composto define um deslocamento a partir da data atual. Por exemplo, `last 10 months (últimos 10 meses)`. Um intervalo composto é formado por uma palavra reservada (`word`), uma granularidade (`granularity`) e opcionalmente por um valor (`value`). A palavra reservada define a direção do deslocamento. A granularidade define a unidade de tempo utilizada no deslocamento. O valor especifica o tamanho do deslocamento. Por exemplo, `last` é a palavra reservada, `10` é o valor e `months` é a granularidade. Se o valor for omitido, um valor *default* para a palavra reservada é assumido. Alguns exemplos de consultas são: (i) `senator Brazil` **`after`** *`2008`*; (ii) `senator Brazil` **`intersect`** *`[01/01/2002, 30/06/2005]`*; e (iii) `senator Brazil` **`before`** *`last 4 years`*.

2.2. Indexação

A ferramenta **Chronos** utiliza dois índices criados a partir das páginas Web coletadas por um *crawler*. Ambos os índices são criados *offline* após a coleta dos documentos. De tempos em tempos, após cada coleta, os índices são atualizados. Durante a fase de consultas, estes índices são acessados pelo mecanismo de busca. O primeiro é um índice invertido convencional construído de acordo com técnicas tradicionais de RI [Baeza-Yates e Ribeiro-Neto 1999]. O segundo é um índice temporal construído em 6 etapas, como observado na Figura 1.

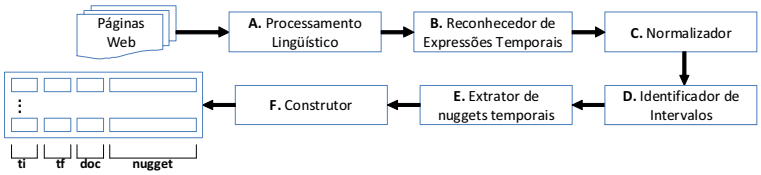

Figura 1. Detalhamento da Indexação Temporal.

A etapa **A. Processamento Linguístico** rotula as palavras do texto de cada página Web com suas funções morfossintáticas, como verbo e substantivo. Esta etapa é realizada pela ferramenta TreeTagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger). A etapa **B. Reconhecimento de Expressões Temporais** anota as expressões temporais presentes no texto. A etapa **C. Normalizador** normaliza as expressões temporais para um formato único. As etapas **B** e **C** são executadas pela ferramenta GUTime

(http://www.timeml.org/site/tarsqi/modules/gutime/index.html), que detecta e normaliza expressões temporais em textos por meio de um autômato finito, considerando as funções morfossintáticas das palavras identificadas na etapa **A**.

A etapa **D. Identificador de Intervalos** identifica quando duas expressões temporais representam um intervalo. Esta etapa é realizada a partir de alguns padrões lexico-sintáticos, tais como: "`from DATE to DATE`" e "`between DATE and DATE`". A etapa **E. Extrator de Nuggets Temporais** extrai *nuggets* do texto que estão associados com expressões temporais. Um *nugget* é um fragmento de uma sentença que informa fatos de domínio aberto associados com alguma entidade [Pasca 2008]. Por exemplo, o fragmento "`O Brasil ganhou a Copa do Mundo em 1994`" é um *nugget* temporal, pois informa um fato (`ganhou a Copa do Mundo`) da entidade `Brasil` e este fato está associado a sua data de ocorrência (`1994`). Esta etapa é baseada nos padrões lexico-sintáticos definidos em [Pasca 2008] e considera as funções morfossintáticas das palavras identificadas na etapa **A**. Um exemplo de padrão utilizado é "$[StartOfSent]\ Nugget\ [in|on]\ Date\ [EndOfSent]$".

A etapa **F. Construtor** é responsável por criar o índice. Cada *nugget* temporal é considerado como um pseudo-documento que é indexado. Alguns campos adicionais são associados a cada *nugget*: (i) o tempo inicial (`ti`) do intervalo representado pela expressão temporal associada ao *nugget*; (ii) o tempo final (`tf`) do intervalo; e (iii) uma referência ao documento que contém o *nugget*. Quando a expressão temporal associada ao *nugget* não constitui um intervalo, e sim um ponto no tempo ("`... elegeu-se em 1994`", por exemplo), atribui-se seu valor tanto para o tempo inicial quanto para o final.

2.3. *Ranking* das Páginas Web

O módulo de *Ranking* das Páginas Web recebe como entrada uma consulta temporal e retorna os resultados. Estes resultados são ordenados de acordo com uma média ponderada entre a relevância do texto, a relevância do tempo e a importância da página. O *score* de relevância de uma página é computado de acordo com a seguinte fórmula:

$$rank(P) = \alpha * ksim + \beta * tsim + \gamma * psim \quad where\ (0 < \alpha, \beta, \gamma < 1)\ \alpha + \beta + \gamma = 1 \quad (1)$$

onde $ksim$ representa o *score* de relevância do texto, $tsim$ representa o *score* de relevância do tempo e $psim$ representa o *score* de importância da página Web. Os valores de α, β e γ representam os pesos das variáveis. O valor de $ksim$ é obtido por técnicas tradicionais do modelo vetorial. Nesta ferramenta usa-se o *score* segundo o critério de relevância do Lucene para a consulta sem a parte temporal sobre o `índice convencional`.

Para computar o valor de $tsim$, primeiramente calcula-se o *score* sobre o `índice temporal`segundo o critério de relevância do Lucene para a consulta sem a parte temporal. Com isso, cada *nugget* temporal indexado recebe um *score*. Em seguida, cada *nugget* temporal cujo intervalo temporal associado não satisfaz o predicado temporal da consulta é eliminado. Finalmente, o *score* da página será o *score* de seu *nugget* temporal com maior *score*.

O valor de β atribui um peso maior se a restrição temporal da consulta é bem restrita, variando entre `0,3` para restrições temporais com intervalos maiores que 30 dias e `0,6` para restrições temporais com intervalos de 1 dia. O valor de α e γ é atribuído

da seguinte forma: $\alpha = \gamma = (1 - \beta)/2$. O valor de $psim$ pode ser obtido de acordo com o algoritmo PageRank [Page et al. 1998], que utiliza a estrutura de *hiperlink* entre as páginas para atribuir sua importância.

Todas as fórmulas acima foram definidas de acordo com [Jin et al. 2008], exceto o cálculo de $tsim$, uma vez que em [Jin et al. 2008] não são utilizados os *nuggets* temporais.

2.4. Apresentação da Ferramenta

A interface da ferramenta **Chronos** é Web, e implementada em Java. Como a maioria dos motores de busca, **Chronos** possui uma interface simples para a realização das consultas. A Figura 2 apresenta a página de resultados. Como resumo de cada documento, nos resultados é apresentado o *nugget* temporal com maior valor de score. Assim, o usuário pode encontrar a informação que necessita já nos resumos apresentados.

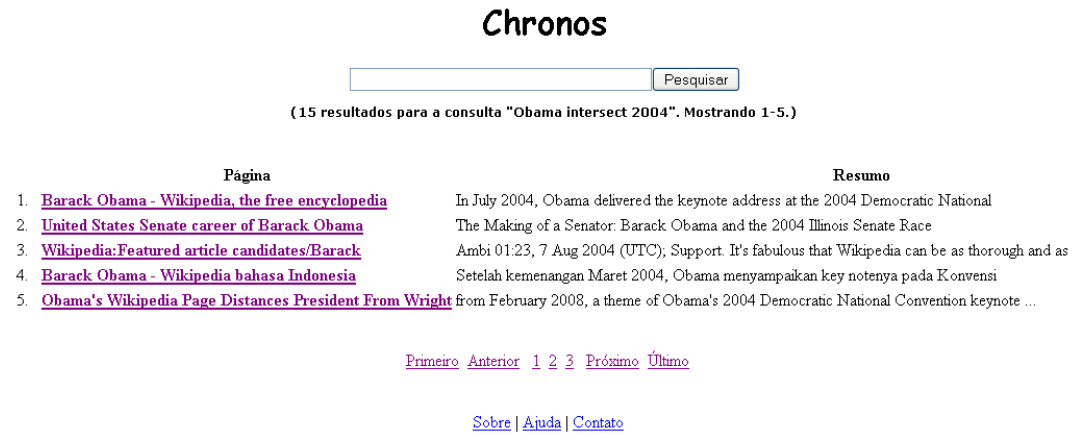

Figura 2. Tela da Ferramenta Chronos.

3. Experimentos

O objetivo destes experimentos é avaliar a qualidade dos resultados retornados pela ferramenta **Chronos**. O conjunto de dados utilizado para realizar a avaliação da ferramenta **Chronos** é composto por 43 páginas coletadas da Wikipedia (http://en.wikipedia.org), com informações sobre os presidentes dos Estados Unidos. A ferramenta **Chronos** foi comparada com o motor de busca convencional Lucene. Foram submetidas 15 consultas temporais aos dois motores de busca, considerando apenas os 10 primeiros resultados e utilizando a mesma importância para todas as páginas ($psim = 1$). Para encontrar as páginas relevantes foram analisados manualmente os resultados de ambas as ferramentas, logo algumas páginas relevantes podem ter sido omitidas. No entanto, como as duas ferramentas não as recuperaram, este fato não prejudica a comparação das ferramentas.

Foram utilizadas duas métricas de qualidade: média das precisões médias (*mean average precision*) e precisão interpolada [Baeza-Yates e Ribeiro-Neto 1999]. A média das precisões médias para **Chronos** foi `0,93` e para Lucene foi `0,49`. Pôde ser observado pelos resultados que a ferramenta **Chronos** obteve um aumento significativo na precisão e também apresentou os resultados relevantes mais ao topo. Isto ocorreu porque a ferramenta **Chronos** normaliza as expressões temporais presentes no conteúdo das páginas e associa a cada expressão temporal apenas os termos relacionados com elas.

No entanto, existem alguns casos em que a ferramenta **Chronos** precisa ser melhorada. Por exemplo, para a consulta "`was assassinated intersect 1881`", havia um documento que continha "`... was mortally wounded in 1881`" que não foi recuperado mesmo contendo a informação desejada. Para resolver este problema, pretende-se utilizar expansão de consultas.

4. Tabalhos Relacionados

TISE [Jin et al. 2008], TERN [Vicente-Díez e Martínez 2009] e [Pasca 2008] são propostas de motores de busca temporal. TISE indexa apenas uma expressão temporal por página, selecionando a expressão temporal que descreve mais apropriadamente os eventos da página Web. Isto acarreta em perda de várias informações temporais presentes no conteúdo da página. TERN e [Pasca 2008] indexam todas as expressões temporais do conteúdo textual da página Web. No entanto, TERN associa todas as expressões temporais com a página, enquanto que [Pasca 2008] associa cada expressão temporal com um *nugget* temporal, formando um pseudo-documento. TISE [Jin et al. 2008] e TERN [Vicente-Díez e Martínez 2009] permitem consultas de seleção temporal, onde o usuário utiliza uma restrição temporal para filtrar os resultados da consulta. Já [Pasca 2008] permite consultas de saída temporal, onde o usuário está interessado em saber qual o tempo em que um determinado evento ocorreu. O diferencial da ferramenta **Chronos** é aplicar a indexação de *nuggets* temporais para consultas de seleção temporal, a fim de melhorar os resultados deste tipo de consulta associando a uma expressão temporal apenas termos que referem-se a ela.

5. Considerações Finais

Este artigo apresentou a ferramenta **Chronos**, que permite a realização de consultas temporais por palavras-chave em páginas Web. As principais contribuições deste trabalho são a identificação dos termos relacionados com cada expressão temporal presente no conteúdo da página Web e a criação de um método de indexação temporal. Experimentos iniciais mostram que **Chronos** melhora a qualidade dos resultados para consultas temporais em comparação com um motor de busca convencional. Como trabalhos futuros destacam-se: (i) a realização de testes com diversos usuários e inúmeras consultas para validar a ferramenta; (ii) a utilização da estrutura da página Web para melhorar os resultados; e (iii) a utilização de expansão de consultas para tratar sinônimos.

Referências

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11).

Baeza-Yates, R. A. e Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Jin, P.; Lian, J.; Zhao, X. e Wan, S. (2008). Tise: A temporal search engine for web contents. In *IITA '08*, pages 220–224, Washington, USA. IEEE Computer Society.

Manica, E.; Dorneles, C. F. e Galante, R. (2010). Supporting temporal queries on xml keyword search engines. In *SBBD '10*, Belo Horizonte, MG, Brasil. SBC.

Page, L. et al. (1998). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

Pasca, M. (2008). Towards temporal web search. In *SAC*, pages 1117–1121. ACM.

Vicente-Díez, M. T. e Martínez, P. (2009). Temporal semantics extraction for improving web search. In *DEXA Workshops*, pages 69–73. IEEE Computer Society.