Proposta de uma Plataforma para Extração e Sumarização Automática de Informações em Ambiente Web

Carlos N. Silla Jr.*, Andre G. Hochuli*, Celso A. A. Kaestner

Pontifícia Universidade Católica do Paraná
Rua Imaculada Conceição 1155, 80215-901 Curitiba - PR - Brasil

{silla,hochuli,kaestner}@ppgia.pucpr.br

Resumo. *Neste trabalho é apresentada a arquitetura de uma plataforma automatizada para a extração de informações e sumarização de notícias em português, obtidas a partir da web. Esta plataforma foi aplicada a um estudo para a extração de informações sobre notícias de jogos de futebol. Foram realizados três experimentos visando estabelecer limites inferiores (baselines) para experimentos futuros, e também para observar o comportamento de sumarizadores existentes atuando de forma complementar à tarefa de extração.*

Abstract. *In this work we present the architecture of an automatic framework for information extraction and summarization of Brazilian Portuguese news, obtained from the web. This framework was applied to a case study for information extraction of news about soccer games. We performed three experiments with the goal of establishing baseline for future experiments and also to observe the behavior of the existing summarizers acting in a complementary way in the task of information extraction.*

1. Introdução

Com a explosão da Internet uma imensa massa de dados oriundos de diferentes fontes passou a se tornar disponível on-line. Um estudo recente de Berkeley [Lyman and H.R., 2003] mostra que em 2002 havia 5 milhões de terabytes de novas informações criadas em documentos impressos, filmes, mídias ópticas e magnéticas. A www sozinha contem cerca de 170 terabytes de informação, o que é equivalente a 17 vezes a coleção da biblioteca do congresso americano. Isto abriu aos usuários a oportunidade de se beneficiar destes dados sob muitas formas [Brin et al., 1998]. Em geral os usuários recuperam informações da Internet por meio de navegação nas páginas (*browsing*) ou pela busca direta de palavras-chave com auxílio de uma máquina de busca [Baeza-Yates and Ribeiro-Neto, 1999].

Entretanto estas estratégias de busca apresentam sérias limitações: (1) o processo de *browsing* não é adequado para a procura de itens de dados específicos, porque o seguimento de links é tedioso e facilmente o objetivo da busca é perdido; (2) a busca por palavras-chave, embora às vezes mais eficiente, retorna em geral grandes quantidades de dados, ultrapassando em muito as condições de manuseio do usuário.

Desta forma, apesar da sua disponibilidade e atualidade, os dados na Internet ainda não podem ser manipulados com tanta facilidade quanto às informações contidas em um Banco de Dados tradicional. Uma possível abordagem para o problema é a extração de dados das fontes disponíveis e sua transferência para uma representação estruturada, no

*Bolsistas PIBIC - CNPq/PUCPR

processo usualmente conhecido como Extração de Informações (*Information Extraction*) [Grishman, 1997].

As aplicações dos sistemas de Extração de Informações são muito variadas; como exemplo podem ser citadas: a obtenção dos autores e da data de apresentação de seminários [Freitag, 1998], a identificação de dados gerais sobre requisitos para empregados em uma base news de anúncios [Califf, 1998], a extração de informações financeiras a partir de sites Internet [Ciravegna, 2001], e a construção de um portal com sumarização e clipping das notícias mais importantes disponíveis nos sites de agências de notícias [McKeown et al., 2003]. Neste último caso também está envolvida a utilização de ferramentas para a sumarização de textos.

O processo de sumarização de textos envolve a construção de uma representação resumida do texto original - um sumário - que preserve as informações constantes no documento original, de acordo com as necessidades do usuário [Luhn, 1958]. Este assunto tem sido objeto de muitas pesquisas recentes [Sparck-Jones, 1999], [Mani, 2001], [Larocca Neto et al., 2002], [Pardo et al., 2003], [Silla Jr. et al., 2003], onde são empregadas técnicas oriundas do processamento de linguagem natural, da aprendizagem de máquina, e da análise e modelagem dos documentos. Pode-se dizer assim que a sumarização de textos atua de forma complementar à tarefa de Extração de Informações.

Neste trabalho é apresentada a arquitetura de uma plataforma automatizada para a extração de informações e sumarização de documentos. Essas tarefas podem ser realizadas independentemente uma da outra, ou pode-se utilizar a sumarização para, inicialmente, reduzir o tamanho do documento, e em seguida ser realizada a tarefa de extração de informações. A área de aplicação definida para os testes é a de extração e sumarização de notícias de futebol. Também são apresentados os resultados obtidos até o momento com a aplicação do sistema.

Este artigo está organizado da seguinte forma: a seção 2 apresenta a arquitetura proposta para o sistema; a seção 3 apresenta a metodologia dos testes efetuados e os resultados obtidos; e na seção 4 discutem-se as principais conclusões do trabalho e indicam-se as próximas etapas que serão realizadas no projeto.

2. A Arquitetura do Sistema

Uma visão geral da arquitetura da plataforma pode ser vista na Figura 1.

Inicialmente um robô de busca é utilizado para a coleta de páginas em sites pré-selecionados, neste trabalho são utilizados sites que contem notícias do campeonato paranense de futebol de 2004. Todas as notícias recuperadas são armazenadas em um banco de dados. Em seguida aplica-se um filtro para selecionar as notícias relevantes. Por exemplo, no contexto desta aplicação existem notícias extraídas da web que falam sobre outros assuntos envolvendo os times que participam do campeonato - como, por exemplo, a saída de um certo jogador para outro time ou futuros confrontos do time em questão - e que não envolvem partidas de futebol.

As notícias consideradas como relevantes que na aplicação descrevem partidas de futebol, vão passar por um pré-processamento, aonde serão extraídos os rótulos HTML. Com a coleção de documentos, pode-se utilizar tanto o processo de sumarização, como o de extração de informações. No caso da sumarização será gerado um sumário contendo as principais informações do texto, e este sumário pode ser apresentado como entrada para o extrator de informações. Já o extrator de informações é responsável por preencher um *template* pré-definido com as informações desejadas; no caso desejam-se informações sobre a partida em questão.

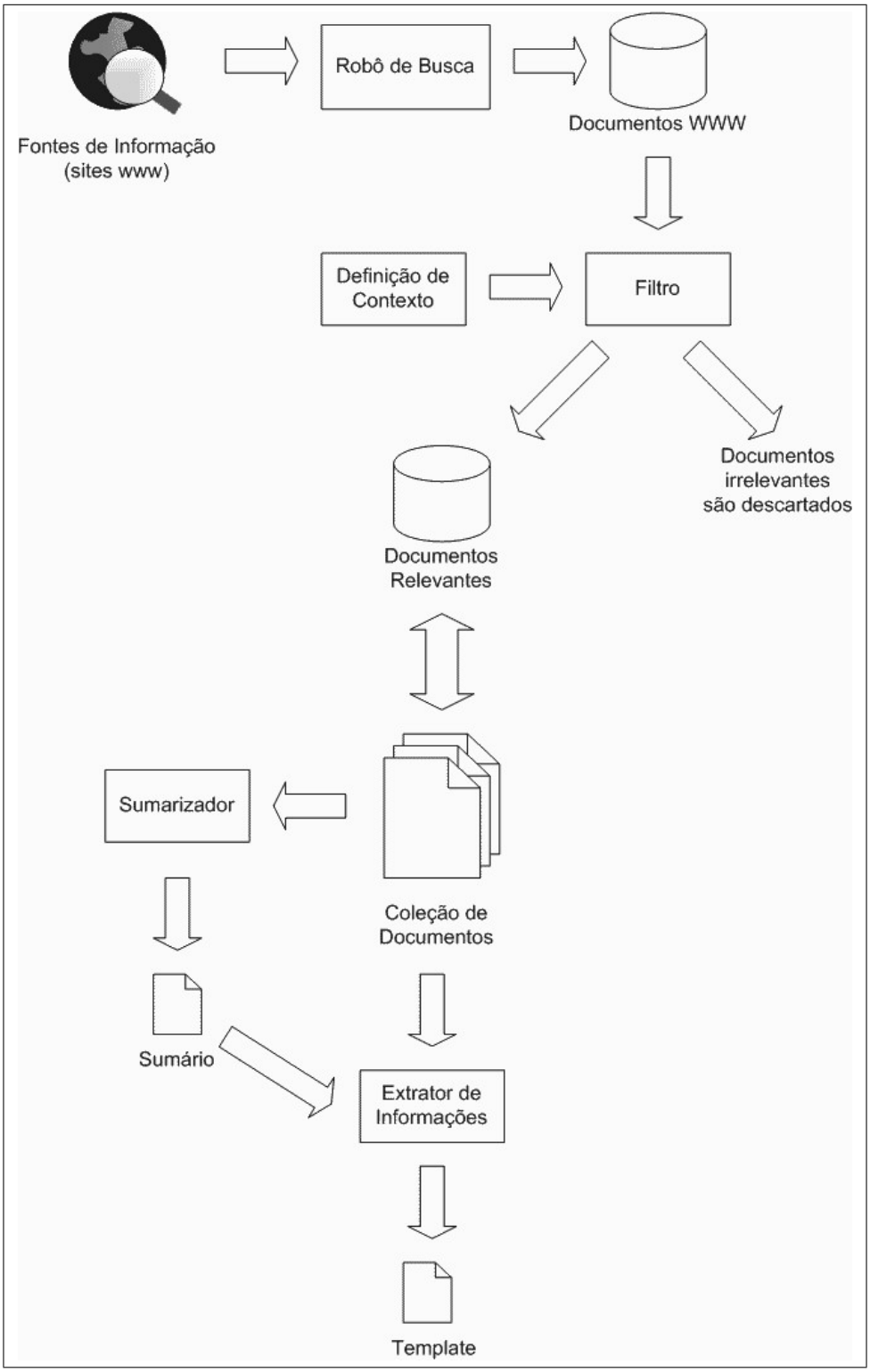

Figura 1: Visão Geral do Sistema

Para a recuperação de notícias dos sites, o robô de busca sendo utilizado é o WebSphinx[1], que foi desenvolvido em java e está muito bem documentado. Para cada fonte de notícia é customizado um robô específico, devido a grande diversidade na estrutura dos sites de notícias desse gênero.

A lógica utilizada pelo filtro é a seguinte: o filtro possui uma lista com todos os times do campeonato e seus possíveis apelidos, como por exemplo, São Paulo e Tricolor. Então, se a notícia em seu conteúdo possuir o nome de pelo menos dois times e também alguma forma de placar, como: [Digito a Digito] (3 a 1); [Digito x Digito] (1 x 0); ela é considerada como relevante.

Os documentos relevantes selecionados foram utilizados para o preenchimento de um *template*, cujas principais informações são: Time1, Time2, Placar. O preenchimento dos *templates* foi manual, por inspeção completa na base.

Em seguida aplicaram-se diversos sumarizadores (ver adiante) aos documentos a fim de verificar se o sumário gerado preserva a informação relevante para o preenchimento do *template*. Tanto o procedimento de sumarização quanto de extração de informações podem ser implementados usando algoritmos de aprendizado de máquina. No estado atual do projeto o método extração de informações usando algoritmos de aprendizado de máquina ainda está sendo implementado e por isso no momento o extrator está utilizado apenas um procedimento simples. O objetivo deste procedimento é o de se obter um valor de limiar mínimo (*baseline)* para os experimentos a serem realizados no futuro, visto que ele preenche o *template* com o time1 e o time2 com os times que mais apareceram na notícia respectivamente. Para isso o procedimento possui uma lista similar à existente no filtro para associar nomes de times e seus respectivos apelidos.

3. Testes Efetuados e Resultados Obtidos

Nesta seção são apresentados os resultados de três experimentos realizados com os seguintes objetivos: (1) verificar a performance do filtro; (2) verificar o acerto da identificação das informações consideradas indispensáveis: a identificação dos times pelo filtro; esta verificação visa utilizar estes resultados como *baseline* para os outros métodos que estão sendo implementados; (3) utilizar os sumarizadores de documentos anteriormente desenvolvidos no intuito de selecionar as sentenças mais relevantes dos documentos. Neste trabalho foram utilizados o FirstSentences Summarizer [Luhn, 1958], TF-ISF-Summ [Larocca Neto et al., 2000] e o ClassSumm [Larocca Neto et al., 2002].

Para a execução do projeto contruiu-se uma base de informações extraída diretamente da www e é composta de 1.169 notícias de sites que continham informações sobre o campeonato paranaense de futebol de 2004. A tabela 1 apresenta a matriz de confusão das notícias selecionadas pelo filtro. Dessa forma a Precisão do filtro na base é de 53,84% e o Recobrimento de 100%. O alto valor de recobrimento é devido a lógica usada pelo filtro, e mesmo tendo sua precisão em 53,84%, esse valor já reduz significativamente o número de notícias que deixaram de ser processadas. No intuito de verificar qual a real eficiência do filtro, são calculadas a micro e a macro médias, sendo que na micro-média considera-se a porcentagem total de acertos, e na macro-média considera-se inicialmente a porcentagem total de acertos por classe, para depois se efetuar a média dos valores obtidos para as classes. Conforme [Manning and Schutze, 2001], no primeiro caso procura-se ponderar a média por cada exemplo, enquanto que no segundo caso busca-se considerar eqüitativamente cada classe. Os resultados obtidos são: micro média de 93,33% e macro média de 96,38%.

[1]Disponível em: http://www-2.cs.cmu.edu/∼rcm/websphinx/

Tabela 1: Matriz de Confusão do Filtro

	Notícia Relevante	Notícia Não-Relevante
Relevante	91	78
Não-Relevante	0	1000

Como visto na seção 2, o extrator utiliza uma heurística para localizar quais foram os dois clubes que jogaram; apesar da tarefa parecer simples muitas vezes não o é, pois constatou-se que existem notícias na base que apresentam até mesmo seis times em seu conteúdo, comentando aspectos do jogo anterior ou do próximo confronto. Porém para se estabelecer um *baseline* para os experimentos futuros, foram verificadas as seguintes taxas de acerto: (1) somente com o nome do 1º time que jogou; (2) somente com o nome do 2º time que jogou; (3) com os nomes do primeiro e segundo times corretos; e por fim (4) os casos onde o nome dos times estivesse invertido. A tabela 2 apresenta os resultados obtidos, mostrando que mesmo uma técnica simples, que foi utilizada para estabelecer um *baseline*, obteve resultados de 48,51% de acerto para a identificação dos nomes dos times.

O terceiro experimento realizado procura verificar quão eficientes são alguns dos sumarizadores desenvolvidos anteriormente para atuarem de forma complementar a tarefa de extração de informação. Ou seja, qual a validade de se usar um sumarizador para tentar inicialmente selecionar as sentenças que contém a informação a ser extraída. Neste contexto um sumário ideal é aquele que possibilita preencher o *template* corretamente. Das notícias que compõem a base verificou-se que a maior parte delas possuem todas essas informações em uma única sentença; porém isso não acontece em todos os casos, podendo essa informação ser encontrada em duas ou até mesmo três sentenças.

Foram utilizados nos experimentos três sumarizadores: (1) FirstSentences que é um sumarizador normalmente utilizado como *baseline* em vários experimentos na área de sumarização; (2) TF-ISF-Summ (Term Frequency - Inverse Sentence Frequency Summarizer) que utiliza uma métrica adaptada do TF-IDF (Term Frequency - Inverse Document Frequency)[Salton et al., 1996] onde a noção de documento é substituída pela noção de sentenças; (3) ClassSumm que utiliza uma abordagem baseada em aprendizado de máquina, sendo que neste trabalho foi utilizado o algoritmo Naïve Bayes. Para realizar os experimentos, devido a necessidade de treinamento do ClassSumm, foi utilizado o método de validação cruzada com fator 10 (*ten-fold cross-validation*) [Mitchell, 1997].

Devido ao enfoque que está sendo utilizado para os testes, não basta comparar os resultados em termos de precisão e cobertura, e sim verificar se o resumo em questão permite preencher corretamente o *template* ou não. Foi estabelecido que, de cada sumarizador seriam escolhidas três sentenças, e que no caso do TF-ISF-Summ e ClassSumm, as sentenças apareceriam por ordem de importância e não pela ordem que estavam no texto (como é comumente utilizado nos experimentos da área de sumarização).

Dessa forma foi utilizado o seguinte método para comparar os três sumarizadores,

Tabela 2: Acerto do Extrator para Detectar os Times (%)

Critério	Número de casos corretos
Time 1	62,37
Time 2	48,51
Time 1 & Time 2	48,51
Time 1 & Time 2 invertidos	28,71

as tabelas 3, 4, 5 apresentam a porcentagem de acerto das sentenças selecionadas em relação a informação desejada, ou seja, quantas vezes cada sentença do sumário gerado contém o time1, o time2 e o placar.

Tabela 3: Acerto do Sumarizador First Sentences (%)

Método: First Sentences	Time1	Time2	Placar
Sentença 1	83,52	70,33	52,75
Sentença 2	14,29	24,18	34,07
Sentença 3	1,10	4,40	8,79
Total de Acerto	98,90	98,90	95,60
Não contém a informação	1,10	1,10	4,40

Tabela 4: Acerto do Sumarizador TF-ISF (%)

Método: TF-ISF-Summ	Time1	Time2	Placar
Sentença 1	70,33	59,34	32,97
Sentença 2	15,38	14,29	14,29
Sentença 3	9,89	10,99	12,09
Total de Acerto	95,60	84,62	59,34
Não contém a informação	4,40	15,38	40,66

Tabela 5: Acerto do Sumarizador ClassSumm (%)

Método: ClassSumm	Time1	Time2	Placar
Sentença 1	75,67	65,78	50,67
Sentença 2	18,78	22,11	28,67
Sentença 3	3,33	4,44	8,78
Total de Acerto	97,78	92,33	88,11
Não contém a informação	2,22	7,67	11,89

4. Conclusões e Direções Futuras

Os resultados apresentados neste trabalho são preliminares visto que o projeto ainda está em andamento. Verificou-se que realizar a etapa de sumarização antes da extração de informações auxilia significativamente o trabalho do extrator, uma vez que este terá que analisar apenas algumas sentenças para preencher o *template* ao invés de todo o documento.

No contexto desta aplicação, o sumarizador que possui o melhor desempenho é o FirstSencentes uma vez que este obteve um acerto de 98,90% para time1 e time2 e um erro de 4,40% para selecionar o placar. Enquanto que o TF-ISF-Summ obteve 95,60%; 84,62%; 59,34% e o ClassSumm 97,78%; 92,33%; 88,11%; de acertos para time1, time2 e placar respectivamente.

Conclui-se para que este tipo de notícia o uso das primeiras sentenças é indicado. Este resultado está em conformidade com a conjectura de que os resultados de um sumarizador dependem fundamentalmente da sua área de aplicação.

Como trabalho futuro será desenvolvido um sumarizador específico para o domínio que está sendo trabalhado, será implementado um extrator de informações utilizando algoritmos de aprendizado de máquina e serão realizados experimentos em outras áreas de aplicação.

Referências

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley.

Brin, S., Motwani, R., Page, L., and Winograd, T. (1998). What can you do with a web in your pocket? *Data Engineering Bulletin*, 21(2):37–47.

Califf, M. E. (1998). Relational learning techniques for natural language extraction. Technical Report AI98-276, Univ. of Texas at Austion.

Ciravegna, F. (2001). Adaptive information extraction form text by rule induction and generalization. In *Proceedings of the 17 th. International Joint Conference on Artificial Intelligence, IJCAI'01*.

Freitag, D. (1998). Information extraction from HTML: Application of a general learning approach. In *Proceedings of the 15th. Conference on Artificial Intelligence, AAAI-98*, pages 517–523.

Grishman, R. (1997). Information extraction: Techniques and challenges. In *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, pages 10–27.

Larocca Neto, J., Freitas, A. A., and Kaestner, C. A. A. (2002). Automatic text summarization using a machine learning approach. In *XVI Brazilian Symposium on Artificial Intelligence*, number 2057 in Lecture Notes in Computer Science, pages 205–215, Porto de Galinhas, PE, Brazil.

Larocca Neto, J., Santos, A. D., Kaestner, C. A. A., and Freitas, A. A. (2000). Document clustering and text summarization. In *Proc. 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining (PADD-2000)*, pages 41–55, London: The Practical Application Company.

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(92):159–165.

Lyman, P. and H.R., V. (2003). How much information. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 [Acesso em: 01/19/04].

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Company.

Manning, C. D. and Schutze, H. (2001). *Foundations of Statistical Natural Language Processing*. The MIT Press.

McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J. L., Nenkova, A., Sable, C., Schiffman, B., and Sigelman, S. (2003). Projeto columbia newsblaster.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Pardo, T. A. S., Rino, L. H. M., and Nunes, M. G. V. (2003). Gistsumm: A summarization tool based on a new extractive method. In *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*, number 2721 in Lecture Notes in Artificial Intelligence, pages 210–218, Germany.

Salton, G., Allan, J., and Singhal, A. (1996). Automatic text decomposition and structuring. *Information Processing and Management*, 32(2):127–138.

Silla Jr., C. N., Kaestner, C. A. A., and Freitas, A. A. (2003). A non-linear topic detection method for text summarization using wordnet. In *1º Workshop em Tecnologia da Informação e Linguagem Humana (TIL)*, São Carlos, SP, Brazil.

Sparck-Jones, K. (1999). *Advances in Automatic Text Summarization*, chapter Automatic Summarizing: factors and directions, pages 1 – 12. MIT Press.