Scientific Automatic Press Observer (SAPO): sistema automático de geração de indicadores de Cultura Científica e de monitoramento de temas científicos na mídia

Carlos Vogt, Flávia Gouveia, Ana Paula Morales, Flávio Daher e Fábio Pisaruk *

Resumo: O Laboratório de Estudos Avançados em Jornalismo (Labjor/Unicamp) vem desenvolvendo nos últimos anos o sistema computacional SAPO *(Scientific Automatic Press Observer),* que coleta, seleciona, organiza e mensura, de forma automatizada, o conteúdo relacionado a temas científicos publicado na mídia *online* não especializada. As matérias extraídas dos veículos analisados são armazenadas em um banco de dados e classificadas por um método que se baseia num conjunto de palavras-chave (*thesaurus*) relacionadas a ciência e tecnologia (C&T). O sistema produz quatro tipos de indicadores de presença de C&T em jornais *online*, contribuindo para o desenvolvimento de indicadores representativos da cultura científica e da percepção pública da ciência. Atualmente, busca-se implementar mudanças e aperfeiçoamentos que tornem o sistema ainda mais consistente e eficaz.

Palavras-chave: comunicação e percepção públicas da ciência, C&T na mídia, indicadores de cultura científica, sistema automatizado de bibliometria, mineração de textos, coleta e classificação automáticas de documentos.

1. Introdução

Informações relevantes encontradas em documentos textuais podem ser – e têm sido crescentemente – identificadas, sistematizadas e utilizadas para subsidiar uma ampla gama de estudos por meio de práticas de mineração de textos. Essas práticas fundamentam-se na organização de bases de dados e procedimentos de classificação e organização de informações, envolvendo sistemas informáticos cada vez mais sofisticados, e resultam em avaliações mais densas e qualificadas quanto mais se apoiem em textos coerentes, confiáveis, bem selecionados e organizados. Os instrumentos já desenvolvidos para tais fins no âmbito de diversos estudos (como os da linguagem, da semiótica, da opinião pública, da sociologia e da antropologia) têm também aplicações na análise da mídia impressa, radiofônica e televisiva (BAUER e GASKELL, 2002).

Fora do contexto acadêmico, verifica-se também o interesse de empresas, instituições, órgãos governamentais e editores de jornais em mensurar sua visibilidade na mídia, avaliar o impacto de políticas na imprensa, monitorar como os efeitos junto ao público leitor evoluem no tempo etc[1]. Para atender a esses diversos interesses, o Laboratório

* Carlos Vogt é coordenador do Labjor/Unicamp e presidente da Fundação Universidade Virtual do Estado de São Paulo Univesp (cvogt@uol.com.br), Flávia Gouveia é doutoranda em Política Científica e Tecnológica pela Unicamp e assessora de comunicação da Univesp (flahgou@uol.com.br), Ana Paula Morales é doutoranda em Política Científica e Tecnológica pela Unicamp e coordenadora de comunicação da Univesp (anapmorales@gmail.com), Flávio Daher é desenvolvedor de infraestrutura de TI do Núcleo de Pesquisas em Políticas Públicas da USP (flaviodaher@gmail.com) e Fábio Pisaruk é engenheiro de software na empresa Bidu (pisaruk@gmail.com).

[1] "Institutos públicos e privados podem precisar analisar o impacto e a repercussão de seus releases para a imprensa, ou de seus posicionamentos públicos. Editores e administradores de jornais podem precisar de instrumentos quantitativos para comparar suas políticas editoriais com a de outros jornais". (VOGT *et al*., 2006)

de Estudos Avançados em Jornalismo da Universidade Estadual de Campinas (Labjor/Unicamp) vem desenvolvendo um sistema de coleta, seleção, organização e mensuração da presença e do impacto de temas de Ciência e Tecnologia (C&T) na mídia *online*, chamado SAPO, sigla em inglês de *Scientific Automatic Press Observer*. Trata-se de um sistema informático baseado em um banco de dados integrado com indicadores quantitativos, medidos automaticamente. Seu objetivo é avaliar a presença de temas científicos na mídia *online* brasileira, permitindo a realização de estudos relacionados com a percepção pública da ciência e da tecnologia, tais como: i) avaliação e mensuração de tendências na cobertura de diferentes temáticas; ii) análise de cobertura de casos midiáticos novos; iii) estudo da evolução temporal de uma notícia e cobertura longitudinal de temas clássicos; iv) estudo de percepção e reposta do público; e v) correlação entre tipo de cobertura de um tema e outras variáveis.

A hipótese que orientou esses esforços é a de que os estudos realizados a partir da mineração de textos em veículos de comunicação, por meio do sistema SAPO, são capazes de monitorar e mensurar a presença de temas da cultura científica na mídia e, assim, refletir a percepção pública da ciência sob uma ótica diferente da tradicional, baseada na aplicação de questionários. O sistema vem incorporando diversos aprimoramentos desde sua concepção e atualmente está estruturado conforme apresenta a próxima seção.

2. O SAPO

2.1 Estrutura e Funcionamento

Para cumprir as funções automatizadas de coleta, seleção, organização e mensuração do conteúdo publicado em veículos *online*, o SAPO armazena esse conteúdo em um banco de dados e o classifica em três categorias. A classificação dos artigos orienta-se por um conjunto de palavras-chave (*thesaurus*), no qual cada palavra-chave possui uma pontuação específica, de acordo com seu peso classificatório. O artigo em análise é aberto para varredura em busca de coincidências e a cada palavra-chave encontrada no artigo adiciona-se o valor de sua pontuação. Ao final da varredura, a pontuação do artigo (*score*) é a soma da pontuação das palavras-chave presentes (contadas apenas uma vez, sem considerar repetições). A pontuação define se o texto é de conteúdo científico; pode ser de conteúdo científico (zona intermediária); ou não é de conteúdo científico.

O sistema se constitui de:

- Um **conjunto de rotinas** preparadas para realizar a varredura do portal de mídia desde 2001 e, em seguida, a indexação e a obtenção de cópia dos textos integrais do portal *online Estadao.com.br*, do jornal brasileiro *O Estado de S.Paulo*.

- Um **sistema de classificação e filtragem** de matérias, capaz de selecionar de forma automática aqueles textos que tratem de temas relacionados a C&T; políticas científico-tecnológicas e de inovação; biomedicina e meio ambiente.

- Um **banco de dados** estruturado e um **buscador inteligente** que organizam e obtêm, a partir de vários metadados (título, caderno, data, fonte, autor), matérias classificadas como “C&T”, “Não C&T” ou “Talvez C&T”. Nesse banco é possível fazer consultas por assunto, por autor, por fonte (agências, assessorias de comunicação institucionais etc.) e por período de tempo (com possibilidade de se realizarem estudos de evolução temporal de notícias sobre um determinado tema).

- Um painel de **indicadores estatísticos**, apresentados de forma gráfica, que permite acompanhar a evolução da frequência, relevância e outras características do material coletado e armazenado.

Os sistemas de varredura, indexação, download, armazenamento e classificação foram desenvolvidos usando-se as linguagens Java[2] e Scala[3], e são agendados para execução automática e cíclica pelo sistema operacional GNU/Linux do servidor. O sistema gerenciador de banco de dados usado para armazenar os artigos, e a partir do qual se processam as consultas de visualização de textos e geração de indicadores, é o PostgreSQL[4]. Para a interface foi usada a linguagem PHP[5] e o framework Smarty[6], em conjunto com um servidor de páginas Apache[7]. A interface web do SAPO é amigável, com opções para busca de palavras-chaves na base de artigos e consulta dos gráficos de indicadores de presença de artigos de C&T na mídia. As solicitações são armazenadas e podem ser analisadas visando melhorias para o sistema.

Assim, todo o processo de alimentação do banco de artigos é automatizado. A intervenção humana somente é necessária para melhorias no filtro do sistema. Uma vez selecionadas, as matérias são arquivadas no banco de dados de forma organizada, acrescentando-lhes um conjunto de metadados. Os filtros permitem identificar as matérias relativas sobretudo às ciências naturais e poderão ser também valiosos na análise de matérias sobre ciências humanas e sociais (item 2.1.1).

2.1.1 Filtro: identificando C&T

Para a elaboração do sistema de filtragem, a opção foi a de considerar a C&T enquanto fenômeno cultural: um grande ecossistema de símbolos, ideias, histórias, fatos, noções, que circulam e inquietam a sociedade e têm, portanto, um forte reflexo midiático. Os mecanismos de seleção do sistema foram organizados de forma a escolher matérias de ciência e tecnologia[8]. As matérias selecionadas são relativas a avanços em tecnologias de ponta ligados a pesquisa (como nanotecnologia, biologia molecular); temas de tecnologia aeroespacial e astronomia; discussões sobre políticas e impacto da CT&I (por ex.: poluição eletromagnética, transgênicos, TV digital); ciências da vida, ciências humanas e sociais, incluindo-se matérias de comportamento ou sobre política e economia que deem voz a pesquisadores dessas áreas, entre outras. Entre as matérias não selecionadas estão os textos sobre produtos tecnológicos (novos modelos de celular, computador etc.) e matérias que apenas expõem dados.

Ao passar pelo filtro, cada matéria adquire uma pontuação (dada pela soma dos pesos de cada palavra-chave encontrada, contabilizados apenas uma vez). As matérias com pontuação igual ou superior a 20 são classificadas como de "C&T". As que têm pontuação igual ou inferior a 17 são classificadas como "Não C&T". Por fim, as matérias situadas na área intermediária (com 18 ou 19 pontos) ficam na categoria "Talvez C&T"[9].

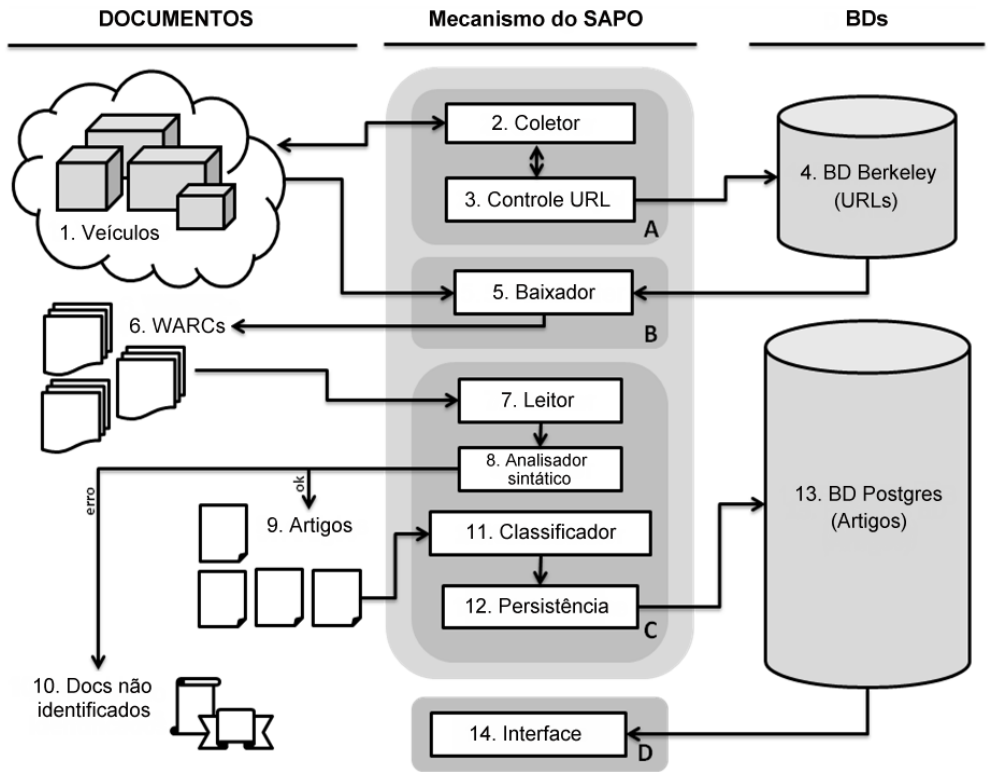
---

[2] http://www.oracle.com/br/technologies/java/index.html
[3] http://www.scala-lang.org/
[4] http://www.postgresql.org/about/
[5] http://www.php.net/
[6] http://www.smarty.net/
[7] http://httpd.apache.org/
[8] O protocolo definidor de C&T encontra-se descrito pela primeira vez no artigo de Vogt *et al*. (2006).
[9] Esses limites de pontuação não são os mesmos empregados na fase inicial do projeto. Os limites eram: para a categoria "C&T", matérias com pontuação superior a 16; para "Não C&T", matérias com pontuação abaixo de 10; e para "Talvez C&T", matérias com pontuação intermediária. Em função dos resultados do teste de confiabilidade (seção 2.2), estabeleceram-se os novos limites de pontuação.

2.1.2 Detalhamentos técnicos

Na concepção e no desenvolvimento do SAPO, priorizou-se a utilização de ferramentas gratuitas e de código aberto, em particular nos seus componentes coletor e baixador. O mecanismo do SAPO é constituído de quatro componentes básicos, identificados com as letras A, B, C e D na Figura 1:

Figura 1: Componentes básicos e funcionamento do SAPO



De forma geral, a construção dos diversos componentes do sistema partiu de programas escritos em linguagens Java e Scala. Todo o SAPO foi desenvolvido a partir de bibliotecas e estruturas integráveis a essas linguagens de programação. A seguir, descreve-se cada componente e sua função no sistema.

2.1.2.1 Coletor e Controlador de URLs

A partir da estrutura do veículo considerado, esse componente (2 e 3, Figura 1) faz um mapeamento de todas as URLs que constituem o universo das notícias e as registra em um Banco de Dados de URLs. O coletor[10] vasculha a web registrando e indexando as URLs que serão posteriormente acessadas pelo baixador. Ele deve ser construído sob medida, tendo em vista a estrutura do portal do veículo que se deseja monitorar. No SAPO, o coletor registra, para cada combinação de dia, mês e ano, as URLs das notícias do dia, por ordem crescente de data. As URLs indexadas passam por um controle automático (3, Figura 1) para evitar duplicidade de registros no Banco de Dados de URLs (4, Figura 1). O controlador de URL faz uma consulta para assegurar que as URLs a serem registradas já não tenham sido incluídas no banco, evitando gasto desnecessário de recursos da máquina. O controlador também monitora a varredura do coletor, detectando a conclusão da indexação de todas as URLs das

[10] Também chamado de crawler, webcrawler, spider, robô ou ainda bot.

notícias de um determinado dia e registrando essa data num campo específico do Banco de Dados de URLs, de modo a criar um ponto de partida para próximas varreduras.

2.1.2.2 Baixador

O mecanismo responsável por obter os documentos indexados pelo coletor é o baixador (B, Figura 1). Ele organiza a fila de requisições e baixa em formato WARC[11] os documentos apontados pelas URLs registradas no Banco de URLs. O sistema (número 5, Figura 1), consiste em: (i) o "baixador" de fato, um programa que faz o download em disco dos documentos apontados pelas URLs levantadas pelo coletor; e (ii) um processo que gerencia e acompanha os trabalhos de download. O processo gerenciador se conecta ao Banco de Dados de URLs e traz para o baixador um conjunto de URLs para iniciar o download, controlando as listas de URLs passadas ao baixador para não haver duplicatas. Outra função do gerenciador é detectar o fim do processo de download da lista de URLs e encaminhar o pacote de documentos recém-obtidos para a etapa seguinte do sistema. O sistema SAPO utiliza o baixador do Heritrix[12], que executa o download dos documentos e os grava em disco no formato WARC. No momento em que o Heritrix termina seu trabalho, o gerenciador libera o arquivo WARC recém-criado para a fase de limpeza e processamento do documento.

2.1.2.3 Limpeza e processamento dos documentos

Esse componente processa os documentos WARC para identificar suas estruturas (título, caderno, conteúdo, data) e os transforma em artigos únicos. Tais artigos são classificados de acordo com seu grau de proximidade com o tema "C&T", e registrados no Banco de Dados de Artigos (13, Figura 1). A fase de limpeza e processamento dos documentos consiste no encadeamento de quatro processos escritos em Java – leitor (7, Figura 1), analisador sintático (8, 9 e 10, Figura 1), classificador (11, Figura 1) e processo de persistência (12, Figura 1) –, que trabalham paralelamente, utilizando um gerenciador de filas desenvolvido especificamente para o SAPO. Cada processo disponibiliza informações, em fila, a serem trabalhadas pelo seguinte. Todos esses processos são gerenciados por um processo-mestre: o Gerenciador de Limpeza.

O leitor abre o arquivo WARC para processamento e separa o conteúdo HTML das páginas. Em seguida obtêm-se os arquivos HTML correspondentes às páginas indexadas inicialmente pelo coletor. Neste ponto, o Gerenciador de Limpeza cria em memória um registro contendo campos para inclusão de informações (URL, data, conteúdo etc.), a ser inserido no Banco de Dados de Artigos quando todos os dados (incluindo a classificação "C&T", "Não C&T e "Talvez C&T") estiverem disponíveis. O arquivo HTML gerado pelo leitor é encaminhado ao analisador sintático, onde um novo processo gerenciador comanda a análise estrutural do documento, visando separar campos no arquivo HTML (URL, data, título, caderno e corpo da notícia). O processo gerenciador aplica uma série de análises no documento HTML[13].

---

[11] Web ARChive, um formato de arquivo comumente utilizado para guardar documentos recuperados por webcrawlers, que combina o conteúdo das páginas HTML com seus cabeçalhos, metadados e demais informações sobre estruturas e verificações de erro.

[12] Aplicativo de código-aberto, escrito em Java e desenvolvido pela biblioteca digital Internet Archive. Esse software arquiva documentos da internet e é utilizado por diversas organizações e bibliotecas nacionais, como a British Library e CiteSeerX. http://en.wikipedia.org/wiki/Heritrix

[13] Caso os campos sejam corretamente identificados, o Gerenciador de Limpeza preenche as lacunas que havia reservado para gerar o artigo (9, Figura 1). Se algum dos campos não puder ser identificado, o documento HTML é considerado não identificado (10, Figura 1) e gravado em uma área do disco. Nesse caso, um registro de erro é escrito no relatório do Gerenciador. O documento não identificado pode ser revisto para verificação do problema e melhoria do processo.

Após o isolamento dos campos do artigo, inicia-se o processo de filtragem para analisar o texto e classificá-lo em “C&T”, “Não C&T” ou “Talvez C&T”. O método classificador busca no conteúdo do artigo cada uma das 597 palavras constituintes do *thesaurus* relativo aos artigos sobre C&T, listadas na tabela sci_filter da base de dados (ver <http://sapo.labjor.unicamp.br/scifilter/>). Cada palavra tem um peso associado. A soma de todos os pesos do artigo, sem repetição, define sua pontuação (*score*). Por fim, o Gerenciador da Limpeza preenche o campo referente ao grupo no qual o artigo foi classificado, estando apto a inseri-lo no Banco de Dados de Artigos.

O processo de persistência realiza a transferência dos dados da memória da máquina para o Banco de Dados de Artigos. Ele controla a escrita no banco, recebendo uma fila de artigos alimentada pelo leitor, e, para cada artigo da fila, compara todos os seus campos com os dos artigos já armazenados no banco, conferindo se há duplicidade. Em caso afirmativo, o artigo é ignorado. Caso contrário, é inserido como um novo registro no banco de dados. É um processo análogo ao Controlador de URLs do coletor.

2.1.2.4 Interface

O SAPO oferece o acesso ao conteúdo coletado e gera indicadores que subsidiam o estudo da divulgação/percepção pública da ciência. Esses recursos são acessados por meio de uma camada específica do sistema, a interface (14, Figura 1), que está disponível aos usuários via internet. Ao acessar o endereço web do projeto (http://sapo.labjor.unicamp.br/), o usuário faz uma requisição de serviço HTTP ao servidor do SAPO. O ambiente de acesso ao SAPO utiliza PHP como linguagem e o Smarty[14] como estrutura básica (framework).

Para apresentar os resultados das buscas, os indicadores e os gráficos, a interface executa consultas (*queries*) predeterminadas ao Banco de Dados de Artigos. Esse banco fornece tabelas e cadeias de textos e números formatadas pelos scripts em PHP para apresentação em formato de página HTML no navegador do usuário.

Quanto às possibilidades de acesso pelo usuário, o SAPO permite realizar: (i) **pesquisas com base em indicadores quantitativos** – o sistema gera gráficos de indicadores por veículo a partir da escolha do status do conteúdo (“C&T”, “Não C&T” e “Talvez C&T”) e do período; (ii) **buscas de conteúdo** – pesquisas qualitativas por termo e recortes temporais; e (iii) **cruzamentos de dados** – estudo quali-quantitativo, combinando indicadores quantitativos e buscas de conteúdo. Nas buscas por conteúdo, podem-se verificar quais temas científicos são objeto de maior interesse para a imprensa, como esses assuntos são tratados e em que seção aparecem mais frequentemente. A busca gera uma página de resultados que dá acesso a metadados das matérias (título, subtítulo, editoria, veículo, data de publicação, total de palavras, pontuação e densidade) e ao texto integral na página do veículo.

2.1.2.5 Bancos de Dados

O SAPO utiliza os bancos de dados Berkeley DB (4, Figura 1), para a relação de URLs, e Postgres SQL (13, Figura 1), para o registro dos documentos armazenados no sistema. O Banco de Dados de URLs abriga as listas de URLs coletadas no módulo

[14] O Smarty provê uma estrutura simples para o grande volume de documentos e as variadas funcionalidades do sistema. Essa estrutura pode ser analisada e atualizada por meio de arquivos .php localizados no diretório raiz do sistema. Para fazer tais análises e atualizações, basta que o corpo técnico tenha proficiência na linguagem PHP.

inicial do sistema. O coletor foi programado de modo a, mesmo se abruptamente finalizado, ser capaz de continuar o processo de onde parou, graças ao uso extensivo de transações[15]. O Banco de Dados de Artigos dá suporte à criação de índices nas tabelas, acelera as consultas e é útil para consultas pré-moldadas, como faz o SAPO ao gerar os indicadores. A tabela com o maior número de registros atualmente no sistema tem mais de 1,1 milhão de linhas, correspondentes aos artigos coletados do *Estadao.com.br*, de janeiro de 2001 a julho de 2013.

2.1.2.6 Veículos

Na atual fase de desenvolvimento do projeto SAPO, optou-se por selecionar o portal do jornal diário *O Estado de S.Paulo*, chamado *Estadao.com.br*. Fundado em 1875, *O Estado de S.Paulo* é um dos maiores jornais em circulação no Brasil. Em março de 2000 ocorreu a fusão dos sites da *Agência Estado*, *O Estado de S.Paulo* e *Jornal da Tarde*, resultando no portal *Estadao.com.br*, veículo informativo em tempo real. O *Estadao.com.br* é um dos portais de notícias de maior audiência da internet brasileira, tendo superado a marca de um milhão de visitantes mensais em janeiro de 2003.

Além da importância do portal para a mídia nacional, outro motivo que levou à eleição do *Estadao.com.br* é a forma como se estrutura seu portal web. Atualmente, o veículo fornece uma lista (arquivo robots.txt acessível em www.estadao.com.br/robots.txt) com links de todas as notícias veiculadas em determinada data, dispostas em índices divididos por dia, mês e ano[16]. Assim, foi possível construir mais facilmente um gerador de URLs para obter os endereços eletrônicos de cada notícia. Outros veículos, como o jornal *Folha de S.Paulo*, *Jornal do Brasil* e jornal *O Globo*, deverão ser incluídos no sistema em nova fase do projeto.

2.1.3 Indicadores

A partir dos dados coletados e classificados, o SAPO possibilita gerar indicadores e gráficos para períodos de tempo específicos. Esses indicadores são instrumentos úteis para auxiliar estudos sobre a mídia *online*, em especial sobre percepção pública da ciência segundo a perspectiva da oferta de matérias científicas pela mídia, como enunciado na hipótese apresentada na introdução deste trabalho, diferente da convencional baseada em questionários aplicados junto ao público.

Para a construção dos indicadores, definiram-se como: N_{tot} o número total de matérias publicadas do veículo de interesse, no período selecionado pelo usuário; P_{tot} o número total de palavras no veículo em análise, em determinado período de tempo; N_{sel} o número de matérias selecionadas pelo sistema, em determinado período para o veículo estudado; e P_{sel} o número total de palavras contidas nas matérias selecionadas pelo SAPO. Os indicadores gerados são:

- **Indicador de massa ($M = N_{sel}$):** número de matérias de C&T no veículo, em determinado período. A análise temporal desse indicador permite evidenciar momentos de "epidemias midiáticas" sobre certos temas científicos, sazonalidades relacionadas a dias da semana em que são veiculadas seções diferenciadas e eventos comemorativos datados. Pode também subsidiar estudos de caso relativos aos temas abordados nos artigos. Calculando-se *M* como uma média sobre intervalos maiores que um dia, consegue-se avaliar o espaço médio dedicado a C&T pelo jornal.

[15] http://en.wikipedia.org/wiki/Database_transaction
[16] Por exemplo, http://www.estadao.com.br/arquivo/2011.

- **Indicador de frequência ($f = M / N_{tot}$):** quantidade relativa de matérias de C&T sobre o total de matérias no veículo, no período selecionado. Esse indicador aponta para o grau de conteúdo científico do veículo relativamente a seu conteúdo total. De forma ainda mais clara que o indicador de massa, revela picos em dias específicos caracterizados pela presença de cadernos intensamente "habitados" por temas de C&T e sinaliza casos midiáticos.

- **Indicador de densidade ($d = P_{sel} / P_{tot}$):** espaço relativo de matérias de C&T, ou seja, porcentagem de palavras dessas matérias sobre o total de palavras no veículo. O indicador *d* foi inspirado nos antigos procedimentos de mensuração da proporção de temas nos jornais impressos, feita com o uso de régua e o cálculo da área ocupada pelo tema de interesse, em centímetros quadrados. O que o SAPO apresenta com este indicador é um resultado semelhante, em porcentagem de palavras dos artigos classificados como "C&T" sobre o total de palavras no veículo, no período em análise. Sua utilidade abrange estudos temporais, comparações temáticas e entre veículos.

- **Indicador de aprofundamento ($A = d / f$):** peso relativo das matérias de C&T em comparação à matéria "média" do veículo. Este indicador combina os indicadores de densidade e frequência. Sendo *A* maior que 1, o veículo está publicando matérias de C&T que são, em média, de tamanho maior que as matérias em geral (matéria média). Observa-se, então, o tipo de política editorial e cultural do jornal. *A* menor que 1 tende a significar uma política editorial que apresenta temas de ciência e tecnologia, de maneira geral, como notícias ou artigos breves.

2.1.4 Resultados

Constantemente monitorado[17] por sua equipe, o SAPO adiciona diariamente conteúdo novo a seus bancos de dados. Mais de 1,1 milhão de artigos do portal *Estadao.com.br* foram coletados pelo sistema desde 1º de janeiro de 2001 (início do período de cobertura) até 31 de julho de 2013. Os resultados apresentados na Tabela 1 referem-se ao período compreendido entre 1º de janeiro de 2001 a 18 de janeiro de 2012. Os artigos classificados como "C&T" compreendem 2,9% do total, com pontuação média de 32,7 e número médio de palavras do *thesaurus* por artigo de aproximadamente 10.

Tabela 1: Resumo de resultados apresentados pelo SAPO por categoria – período: 01/jan de 2001 – 18/jan de 2012

		Categorias		
		C&T	**Talvez C&T**	**Não C&T**
Artigos	Qtd.	30.544	5.002	1.027.471
	%	2,9	0,5	96,7
Pontuação	min	20	18	-12
	média	32,7	18,5	1,5
	max	193	19	17
Palavras-chave por artigo	min	3	3	0
	média	10,3	6,5	0,7
	max	57	14	14

[17] A equipe do SAPO checa periodicamente se o portal coberto pelo sistema não eliminou ou criou seções, mudou a estrutura de arquivamento das matérias etc.

2.1.5 Buscas

Este item traz os tipos de resultados do SAPO para buscas por conteúdo, por meio de exercícios exemplificadores. O primeiro deles tem as seguintes especificações: período de 17 de junho de 2010 a 17 de junho de 2011; nenhum termo selecionado; status “C&T”. O sistema encontra 6.446 artigos de C&T. A página de resultados (Figura 2) apresenta o título com o link de cada artigo selecionado, sua pontuação, a editoria em que foi publicado, a data e o total de palavras.

Figura 2: Página de resultados de busca do SAPO para o período de 17/jun de 2010 a 17/jun de 2011, nenhum termo de busca selecionado e status “C&T” – *Estadao.com.br*

Em um exercício similar, selecionou-se desta vez com o termo “transgênico”, mantendo-se o período. Neste caso, o sistema encontrou 49 artigos de C&T com o termo. Outros exercícios de busca por conteúdo podem ser realizados, alterando-se o status – para, por exemplo, “Talvez C&T” (com 5 resultados no mesmo período, para o termo “transgênico”) ou “Não C&T” (29 resultados no período, termo “transgênico”).

2.1.6 Indicadores por período

Esta seção apresenta exemplos de resultados de uma pesquisa de indicadores no mesmo período selecionado na seção anterior, de 17/jun de 2010 a 17/jun de 2011. O Gráfico 1 mostra os valores assumidos pelo indicador de massa mês a mês e permite, assim, identificar os meses de pico de publicação de artigos de conteúdo científico (setembro de 2010 e março de 2011) e os meses em que os temas de C&T foram menos presentes em termos absolutos (fevereiro de 2011, julho de 2010 e maio de 2011). De posse dessas informações, o usuário do sistema pode realizar uma busca nos meses de seu interesse no campo “Busca”, optando por não especificar termos, e obter uma página de resultados com acesso aos artigos. A leitura dos títulos e, eventualmente, do conteúdo dos artigos selecionados revelará os temas abordados e circunstâncias relacionadas, refinando assim a pesquisa e auxiliando a interpretação de resultados.

Gráfico 1: Indicador de Massa para o período de 17/jun de 2010 a 17/jun de 2011, em valores mensais – veículo: *Estadao.com.br*

Gráfico 2: Indicador de Frequência para o período de 17/jun de 2010 a 17/jun de 2011, em valores mensais – veículo: *Estadao.com.br*

Gráfico 3: Indicador de Densidade para o período de 17/jun de 2010 a 17/jun de 2011, em valores mensais – veículo: *Estadao.com.br*

Gráfico 4: Indicador de Aprofundamento para o período de 17/jun de 2010 a 17/jun de 2011, em valores mensais – veículo: *Estadao.com.br*

O Gráfico 2 traz os percentuais de matérias de C&T em relação ao total de matérias publicadas no portal *Estadao.com.br* no período selecionado e revela que o pico absoluto de matérias de C&T apontado pelo indicador de massa corresponde a um pico da proporção de artigos científicos no período. O indicador de frequência mostra que o pico relativo do período se dá em março de 2011 (valor próximo de 10%).

O próximo indicador calculado para o mesmo período é o de densidade, que mensura o espaço relativo dos artigos de C&T em relação ao total publicado no veículo. Conforme o Gráfico 3, é no mês de outubro de 2010 que o portal *Estadao.com.br* ofereceu maior espaço relativo de publicação para artigos de conteúdo científico. Por outro lado, os meses de menor densidade foram junho de 2010 e fevereiro de 2011.

Finalmente, analisando-se o indicador de aprofundamento, que aponta o peso relativo dado aos artigos de C&T pelo veículo em análise no período, encontra-se o pico (*d* / *f* aproximadamente igual a 12) em maio de 2011 (Gráfico 4). Assim, as matérias de C&T veiculadas nesse mês provavelmente trataram os temas com mais profundidade (mais palavras por artigo) que em outros meses do período considerado. Para qualquer dos indicadores, a busca sem especificação de termos auxilia o detalhamento da pesquisa.

2.1 Avanços na confiabilidade

Desde sua primeira versão, o SAPO vem passando por avaliações que conduzem a novos desenvolvimentos. Em sua fase inicial, o SAPO foi submetido a testes de confiabilidade[18] que indicaram uma margem de erro total dos dados quantitativos fornecidos, para o conjunto das categorias de classificação, não maior que ±3% (VOGT *et al*., 2006). Na análise sobre as matérias de ciências naturais e exatas, a margem de discordância do sistema com codificadores humanos treinados na classificação das matérias de C&T foi de aproximadamente 10%. Para "Não C&T", o erro do sistema foi menor que a discordância entre dois humanos treinados (< 0,5%).

Posteriormente, um novo modelo de testes foi empregado para avaliar o desempenho do sistema e a necessidade de ajustes no processo de classificação. Criou-se um banco de documentos previamente classificados por leitores especializados, segundo os critérios de classificação descritos no item 2.1.1 deste artigo. Para a classificação manual, foram apresentados à equipe de pesquisa, por meio de uma interface de acesso restrito, artigos sorteados de um conjunto previamente extraído do banco do SAPO (o conjunto é composto de artigos publicados entre os dias 1º de janeiro de 2010 e 1º de janeiro de 2011[19]). O pesquisador lia a matéria e então decidia se a notícia se encaixava no tema C&T, clicando em um dos botões apresentados (C&T, Não C&T e Talvez C&T). O resultado era armazenado numa tabela no banco de dados e outra notícia era exibida para o pesquisador.

Para testar a confiabilidade do SAPO, compararam-se os resultados da classificação manual com o resultado da classificação automática do sistema sobre os mesmos documentos. Utilizou-se a técnica da amostragem aleatória estratificada com base na classificação automática prévia. Para a classificação manual, foram selecionados artigos do conjunto publicado entre 1º de janeiro de 2010 e 1º de janeiro de 2011, com

[18] Os testes realizados foram: a) nível de concordância entre humanos e máquina para as matérias descartadas; b) nível de concordância entre humanos e máquina para as matérias selecionadas; c) situação e composição das matérias classificadas pelo sistema como "talvez ciência" e d) *intercoder reliability* entre máquina e humanos (para garantir que o nível de concordância entre a máquina e humanos sobre matérias "de ciência" não seria menor que entre dois codificadores).

[19] A limitação temporal do teste justificou-se por restrições técnicas dos servidores que hospedavam o SAPO. Planeja-se estender o teste para todo o intervalo temporal coberto pelo sistema.

diferentes probabilidades para cada categoria: 2/5 para aqueles classificados como sendo de “C&T”, 2/5 para os “Não C&T” e 1/5 para os artigos “Talvez C&T”. Essas proporções garantem a presença equilibrada de artigos de todas as categorias, eliminando a tendência de se considerar predominantemente os artigos “Não C&T”, que são a maioria da base (mais de 90% dos textos, conforme a Tabela 1).

2.1.1 Avaliação da metodologia de classificação e recalibragem

O procedimento de avaliação baseou-se na classificação de 2.006 artigos, o equivalente a 1,45% do total de artigos coletados e classificados pelo SAPO no período considerado para o teste[20]. Para efeito de classificação do SAPO, as categorias são disjuntas, ou seja, cada artigo pode pertencer a somente uma das três categorias de documentos. A distribuição dos 2.006 artigos entre as categorias de classificação automática (segundo os limites estabelecidos antes da calibragem) e manual é mostrada na Tabela 2.

Tabela 2: Distribuição dos artigos do banco amostral de avaliação

		Classificação Manual						Total Automática	
		C&T		Não C&T		Talvez C&T			
Classif. Automática	C&T	396	20%	172	9%	83	4%	651	32%
	Não-C&T	30	1%	982	49%	31	2%	1.043	52%
	Talvez-C&T	85	4%	191	10%	36	2%	312	16%
Total Manual		**511**	25%	**1.345**	67%	**150**	7%	**2.006**	100%

A análise utilizada foi a da Precisão-Revocação, usual para avaliação do desempenho de classificadores de texto (SEBASTINI, 2002). A classificação ternária foi convertida em três classificadores binários, um para cada categoria i = {C&T, Talvez C&T, Não C&T}. Esses classificadores comparam as avaliações dos especialistas com a classificação automática. Os resultados possíveis do teste de classificação para a categoria i são: VP_i - Verdadeiro-Positivo: resultado correto (pertence à categoria); FP_i - Falso-Positivo (erro tipo I): alarme falso; FN_i - Falso-Negativo (erro tipo II): falha do alarme; e VN_i - Verdadeiro-Negativo: resultado correto (não pertence à categoria).

A avaliação baseou-se no estudo das medidas Precisão, Revocação e *Medida* F_1, descritas no Quadro 1, para determinada categoria i do total de N ($N = 3$ no caso do SAPO). A redefinição dos valores-limite dos intervalos da pontuação para cada categoria foi estabelecida a partir da otimização das medidas de desempenho calculadas sobre a base de avaliação. Os 2.006 artigos da amostra de classificação foram distribuídos conforme sua pontuação e observados em relação aos pontos de corte que delimitavam as categorias do classificador automático antes da calibragem, isto é, com a categoria “Talvez C&T” delimitada pelas pontuações maiores que 9 e menores que 17. Combinando-se o resultado dos procedimentos de calibragem com base nas medidas P, R e F_1, o intervalo de pontuação para o classificador “Talvez C&T” foi alterado para [18, 19]. Definiram-se assim, consequentemente, os pontos de corte para os demais classificadores.

[20] Com base na suposição simplificada e conservadora de que o valor do estimador da proporção de artigos de C&T segue uma distribuição Normal, a amostra de 2.006 artigos utilizada para o teste de confiabilidade implica em um erro amostral de 2,3%, com probabilidade de 95%.

Quadro 1: Medidas utilizadas para a avaliação de confiabilidade do SAPO

Medida	Fórmula	Descrição
Precisão (*P*)	$P_i = \frac{VP_i}{VP_i + FP_i}$	Fração dos documentos classificados simultaneamente pelos especialistas e pelo SAPO como pertencentes à categoria *i* (concordância homem-máquina) dentre todos os artigos atribuídos automaticamente pelo SAPO a essa categoria.
Revocação (*R*)	$R_i = \frac{VP_i}{VP_i + FN_i}$	Fração dos documentos classificados simultaneamente pelos especialistas e pelo SAPO como pertencentes à categoria *i* (concordância homem-máquina) dentre todos os artigos atribuídos pelos especialistas a essa categoria.
Medida F_1	$F_1 = \frac{2 \cdot P \cdot R}{P + R}$	Média harmônica entre *P* e *R*. Pode ser calculada para as categorias individuais *i* e para as medidas que agregam categorias (macromédia[21] e micromédia[22]).

2.1.2 Resultados

As técnicas de avaliação de desempenho aplicadas sobre a base de artigos classificados manualmente permitiram uma eficiente calibragem dos limites de pontuação do sistema de classificação. Os valores das medidas para os classificadores antigos e os novos são apresentados na Tabela 3.

Tabela 3: Medidas de avaliação dos classificadores antigos e novos, sobre a base de classificação humana

		Categorias			Avaliação geral	
Classificadores		**Não C&T**	**Talvez C&T**	**C&T**	**Macro-média**	**Micro-média**
Novos Talvez = [18,19]	*Nº de documentos*	*1.404*	*72*	*530*	-	-
	Precisão	0,853	0,236	0,668	0,586	**0,782**
	Revocação	0,89	0,113	0,693	0,565	**0,782**
	Medida F_1	0,871	0,153	0,68	0,575	**0,782**
Antigos Talvez = [10,16]	*Nº de documentos*	*1.043*	*312*	*651*	-	-
	Precisão	0,942	0,115	0,608	0,555	**0,705**
	Revocação	0,73	0,24	0,775	0,582	**0,705**
	Medida F_1	0,822	0,156	0,682	0,568	**0,705**

A medida mais adequada para a avaliação global do sistema de classificação é a F_1 agregada pela micromédia, de valor 0,782 após a calibragem. Essa medida pode ser interpretada da seguinte forma: 78,2% dos artigos foram corretamente classificados pelo SAPO. Note-se que a calibragem promoveu um aumento da confiabilidade.

[21] Média aritmética da medida de avaliação (*P* ou *R*) entre as *N* categorias (3, no caso do SAPO).

[22] Cálculo similar a *P* e *R*, acumulando-se os resultados do teste de mesma natureza para cada categoria:

$$P^m = \frac{\sum_{i=1}^{N} VP_i}{\sum_{i=1}^{N} (VP_i + FP_i)}, \quad R^m = \frac{\sum_{i=1}^{N} VP_i}{\sum_{i=1}^{N} (VP_i + FN_i)}.$$

Para a classificação baseada em pontuação de palavras-chave, o desempenho geral do sistema atualmente é comparável ao de classificadores que usam técnicas mais sofisticadas, como as probabilísticas e de aprendizado de máquinas (YANG, 1999). Tal resultado contou com os valores mais elevados de precisão e revocação da categoria "Não C&T" e também com as medidas de precisão e revocação do classificador "C&T", que indicam a adequação da lista de palavras-chave, com controle de sinônimos e conceitos semelhantes, e agrupamento de palavras segundo um radical comum.

Essas estatísticas permitem avaliar o classificador e acompanhar sua evolução quando de ajustes específicos, como alterações no *thesaurus*, redimensionamento das pontuações que definem as categorias ou reformas técnicas na coleta. Convém notar que parte dos erros do sistema é intrínseca à definição de sentido amplo dada a C&T. Mesmo durante a construção da base de testes de confiabilidade, 18 artigos (cerca de 0,9%) não foram incluídos por haver discordância das classificações humanas quanto às categorias nas quais seriam enquadrados, restando 2.006. Aprimoramentos deverão aumentar a confiabilidade e a eficiência do SAPO (seção 3).

3. Melhoramentos no SAPO

Como se pôde ver, os resultados preliminares de exercícios de busca utilizando o SAPO estabelecem-no como uma ferramenta útil para estudos midiáticos e consistente em seus aspectos fundamentais, capaz de mensurar a presença de temas científicos na mídia e de fornecer indicadores relacionados.

A partir de discussões realizadas pelos membros da equipe do SAPO e com colegas durante os encontros realizados na London School of Economics and Political Science, na Inglaterra, em 2011 e 2012, e das pesquisas desenvolvidas pelo grupo do Labjor/Unicamp, planeja-se implementar um conjunto de medidas no sentido de aprimorar o sistema em diversos aspectos. O Quadro 2 apresenta uma relação dessas medidas, agrupadas segundo o componente do sistema associado.

4. Conclusões

Os testes de busca no SAPO indicam que o sistema é de grande utilidade para auxiliar estudos sobre a mídia *online*, com indicadores e resultados consistentes. O sistema oferece à comunidade de pesquisadores informações diárias sobre a cobertura da mídia em C&T, permitindo enxergar não só quanto este tema frequenta o leitor, mas também como o leitor o frequenta nos jornais.

O SAPO inspirou desenvolvimentos de sistemas similares por pesquisadores de outros países, como Itália (sob coordenação do Prof. Dr. Federico Neresini, da Universidade de Pádua - grupo PaSTIS[23]), Turquia (coordenação do Prof. Dr. Ahmet Süerdem, do Departamento de Administração de Negócios da Istanbul Bilgi University) e Inglaterra (grupo coordenado pelo Prof. Dr. Martin Bauer, da London School of Economics – Methodology Institute). Nos encontros realizados em 2011 e 2012 na London School of Economics, sob coordenação do Prof. Martin Bauer, o SAPO e outros sistemas análogos foram debatidos. Novos encontros e discussões deverão ocorrer, visando não apenas ao compartilhamento de conhecimentos entre as equipes responsáveis pelos sistemas, mas também, possivelmente, à formação de um grupo internacional de cooperação, com foco em sistemas informatizados de mineração de textos para mensuração da presença de temas científicos em mídias *online*.

[23] Padova Science, Technology & Innovation Studies.

Quadro 2: Aperfeiçoamentos previstos para o SAPO, por componente do sistema.

Infraestrutura	Coleta	Analisador sintático	Classificação	Interface	Confiabilidade
Migração para servidor em nuvem: maior flexibilidade para o desenvolvimento de aplicações web.	**Revisão do coletor**: permitirá montar máquinas virtuais dedicadas somente à indexação de URLs, tornando a coleta mais rápida.	**Interface para identificação de problemas de análise sintática**: usuários poderão localizar problemas de análise sintática em páginas.	**Categorias**: áreas do conhecimento; outras categorias; seção do veículo; agente da cult. científica; atores; repercussão de audiência.	**Opções de ordenação dos resultados das buscas**: segundo critérios como a pontuação e o número de palavras por artigo.	**Revisão da metodologia de avaliação**: novas técnicas serão empregadas para avaliar a confiabilidade do sistema.
Mudança de framework da aplicação web: aplicações serão desenvolvidas em Ruby on Rails, plataforma mais eficiente, segura e prática.	**Inclusão de novos veículos**: será possível gerar novos dados e comparar os veículos.	**Reconhecimento de campos gráficos**: solução baseada no reconhecimento de campos gráficos na página que contém o artigo (título, linha fina, autor etc.).	**Metodologias**: Palavras-chave com pontuação normalizada; *Machine Learning*; Analisadores morfológicos e sintáticos.	**Reforma no layout, formas de busca e gráficos**: melhor navegabilidade, buscas de vários termos e gráficos de indicadores para buscas por termos.	
	Adaptação para outras bases de informação *online*: blogs, *feed*s RSS e redes sociais.			**Mais dados**: incluir bancos de dados relacionados, permitindo cruzar informações.	
	Novas formas de comunicação na mídia *online*: conteúdos de áudio, imagens, tiras etc.			**Exportação de planilhas**: para armazenar dados de pesquisa dos usuários.	

Bibliografia

BAUER, M.W.; GASKELL, G. (2002) *Pesquisa qualitativa com texto, imagem e som. Um manual prático*. Vozes , Petrópolis, RJ.

VOGT, C.A. (2011). *The spiral of scientific culture and cultural well-being: Brazil and Ibero-America*. Public Understanding of Science (Print), v. 1, p. 1-13.

VOGT, C.A.; CASTELFRANCHI, Y.; RIGHETTI, S.; EVANGELISTA, R.A.; MORALES, A.P.; GOUVEIA, F. (2011) Building a science news media barometer SAPO. In: Bauer, M.; Shukla, R.; Allum, N.. (Org.). *The culture of science - how the public relates to science across the globe*. 1ª ed. New York/London: Routledge, p. 400-413.

VOGT, C.A. *et al*. (2006) SAPO (Science Automatic Press Observer): Construindo um barômetro da ciência e tecnologia na mídia. In: *Cultura científica: desafio*s. EDUSP FAPESP, São Paulo, p.85-130.

SEBASTIANI, F. (2002) *Machine learning in automated text categorization*. In: ACM Computing Surveys (CSUR), Volume 34 Issue 1, ACM New York, NY, USA.

YANG, Y. (1999) An Evaluation of Statistical Approaches to Text Categorization, Information Retrieval, v.1 n.1-2, p.69-90. Disponível em <http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA327980>. Acesso em julho de 2013.