O LEAD AUTOMATIZADO: UMA POSSIBILIDADE DE TRATAMENTO DA INFORMAÇÃO PARA O JORNALISMO IMPRESSO DIÁRIO

AUTOMATED LEAD: A POSSIBILITY FOR INFORMATION TREATMENT FOR DAILY PRESS JOURNALISM.

Tacyana Arce
<tacyarce@hotmail.com>

Resumo: *Apresenta-se a automação do lead das notícias como uma possibilidade de aceleração do processo de produção jornalística e de simplificação do processo de tratamento da informação produzida nos jornais impressos diários. Apresentam-se os pressupostos da automação da notícia, baseados em sua estrutura lógica. Evidencia-se que a ausência de sistemas de recuperação da informação dificultam o trabalho dos jornalistas, que precisam, freqüentemente, recorrer a material anteriormente produzido, para fazer um jornalismo mais analítico e crítico. Propõe-se que a mesma base que possa dar origem a uma produção automática da notícia seja usada para automatizar a análise de assunto dessa mesma notícia, abreviando o processo indexação e tornando-o mais compatível com a quantidade de volume e a agilidade do jornal diário.*

Palavras-chave: *Análise de Assunto, Indexação, Sistemas de Recuperação da Informação, jornalismo, linguagem jornalística, notícia, estrutura da notícia, lead, informação, discurso, automação, jornal impresso*

Abstract: *This article presents news lead automation as a possibility to speed the process of journalism production and simplification of the information treatment produced in daily press. The lack of Information Retrieval systems deter the journalist´s work. This articles proposes that the same Data Base used for a more analytical journalism be also used to generate an automatic lead, and subject analysis, expediting the process of indexation.*

Key-Words: *Subject analysis, Indexation, Information Retrieval System, Journalism, automation*

Introdução

A crescente capacidade de memória e velocidade de operação das máquinas, progressos na área de sintaxe e inteligência artificial. Esses são alguns dos fatores que tornam cada vez mais possível, e real, a produção automática de discursos, isto é, o desenvolvimento de programas capazes de transformar automaticamente em informação eventos

observados por algum dispositivo periférico, ou dados digitados em um computador.

É assim, por exemplo, que, sem que alguém tenha o trabalho de ficar somando pontos obtidos por cada time de futebol no Campeonato Brasileiro, pela simples digitação, em uma planilha eletrônica, dos resultados das partidas realizadas semanalmente, sites especializados em coberturas esportivas mantém a tabela de classificação do campeonato atualizada em tempo real, incluindo gráficos e comentários sobre o histórico do desempenho de cada time.

Também é pela geração automática de discursos que em questão de segundos é possível obter, também na internet, relatórios sobre o mapa astral do interessado que concorde em perder alguns minutos fornecendo dados pessoais, como horário e local de nascimento, algo que antes demandava uma consulta esotérica.

Os campos inicialmente imaginados para a aplicação do recurso da automação de discursos eram técnicos e científicos, como observações meteorológicas e registros médicos, mas, atualmente, sua utilidade se estende desde o entretenimento à elaboração de políticas públicas (a versão on-line do Atlas do Desenvolvimento Humano da Fundação João Pinheiro oferece uma análise de 15 páginas sobre o desenvolvimento social e econômico de cada município brasileiro, bastando que o usuário informe quais sub-índices do IDH pretende cruzar).

A produção automática de discursos pressupõe o preenchimento de algum tipo de formulário com variáveis, a ordenação dessas variáveis segundo algum critério (no caso do Atlas mencionado acima, podem ser renda, PIB per capita, esperança de vida ao nascer, escolaridade da população, mortalidade infantil, número de crianças exploradas no trabalho infantil, número de adultos analfabetos etc) e, por fim, a produção sintática do enunciado final.

Vários são os campos científicos que têm trabalhado com a questão da automação do discurso. A Lingüística contemporânea busca o estabelecimento de padrões computáveis para as línguas naturais. A Lógica busca inovações instrumentais para a compreensão de códigos lingüísticos. A Psicologia Cognitiva investiga como cérebros humanos conseguem, a partir de seu conhecimento de mundo, eliminar a ambigüidade da maioria dos enunciados.

A Ciência da Informação tem estudos, ainda que incipientes, de automação da leitura técnica para a determinação dos assuntos dos documentos, uma das etapas do tratamento da informação.

Em julho de 1997, durante o XX Congresso da Intercom, em Santos, Nilson Lage, coordenador do curso de Jornalismo da Universidade Federal de Santa Catarina, apresentou uma possibilidade de automação do discurso jornalístico a partir do *lead*, como é chamado o primeiro parágrafo de uma notícia, uma vez que este, ao menos como previsto na Teoria do Jornalismo, é uma estrutura lógica. Na síntese acadêmica de Laswell *apud* LAGE (2004), o *lead* informa *quem* fez *o que*, *a quem, quando, onde, como, por que* e *para quê*. Ressalta-se que a possibilidade teórica levantada por LAGE (2004) não foi traduzida em experimento, o que também não ocorrerá neste estudo.

Analisa-se a proposta de LAGE (2004) à luz dos pressupostos da Ciência da Informação, especificamente do tratamento da informação[1], entendendo-se que tal abordagem traz benefícios para as duas áreas. Para o Jornalismo, fortalecer a estrutura da notícia e buscar a automação do discurso não apenas acelera a produção de notícias (um dos grandes problemas do jornal diário é a premência do tempo) como favorece a recuperação da informação[2], outro nó do jornalismo (sempre que acontece um incêndio com centenas de mortos, ou a queda de um Boing lotado, é necessário remeter a eventos similares anteriores, mas, na prática, nem sempre é simples recuperar essas informações, armazenadas em bancos de dados).

Para os pesquisadores da Ciência da Informação – campo amplo que se dedica à investigação e análise de fenômenos ligados à produção,

[1] No Jornalismo, a expressão "tratamento da informação" refere-se à maneira como a informação é manipulada para a geração de notícia, ou seja, como (e sob quais ângulos) ela é abordada, quem são as fontes de informação, quais são os critérios (ideológicos, políticos, mercadológicos ou culturais) que definem o que é ou não notícia para determinado veículo e, portanto, para determinado público. Neste artigo, entretanto, a expressão será usada na perspectiva da Ciência da Informação, ou seja, como um processo de seleção, organização, classificação, indexação e disponibilização da informação para posterior uso.

[2] O termo recuperação da informação (information retrieval) foi cunhado por Calvin Mooers em 1951. Refere-se à possibilidade de se acessar a informação necessitada pelo usuário. Para que uma informação possa ser recuperada, é necessário que ela esteja tratada e inserida em um Sistema de Informação, ou Sistema de Recuperação da Informação (SRI), tipo de sistema de comunicação que, entre outras funções, visa a dar acesso às informações nele registradas (ARAÚJO, 1994)

organização, difusão e uso da informação em todos os campos do saber – pode ser de grande ajuda, e importância, entender como se processa a informação que chega até os leitores de um jornal diário. Afinal, se os impressos não têm a mesma abrangência quantitativa que as mídias eletrônicas e digitais – como televisão, rádio e internet – eles ainda têm forte presença sobre a parcela da sociedade formadora de opinião, o que os torna peças fundamentais na discussão sobre a democratização de acesso à informação e da construção da sociedade da informação.

1 – Notícia e linguagem jornalística

1.1 – Estrutura da notícia

Notícia, na definição de LAGE (2004), é o

> "*relato de um fato novo e relevante, tomado a partir de seu aspecto mais importante ou interessante; ou de uma série de fatos novos relevantes, tomados a partir do fato mais importante ou interessante*".

Não se deve confundir notícia e jornalismo. A notícia é uma das possibilidades do jornalismo, que pode se manifestar ainda sob forma de reportagem, notas, fotografias, infografias, colunas de serviço.

Especificamente, para a possibilidade de automação do discursivo informativo levantada pelo autor, notícia deve ser diferenciada de reportagem. A primeira refere-se a um fato (ou conjunto de fatos), a segunda, a um assunto. Noticia-se a morte de sem-terra numa invasão que acaba de ocorrer em determinada região (hora e local, envolvidos, circunstâncias, responsáveis, repercussão). Faz-se uma reportagem sobre o conflito fundiário (panorama histórico, contexto social, político e econômico, ausência ou necessidade de políticas públicas, reflexos na economia e desenvolvimento do País, implicações internacionais etc).

Tal distinção se faz necessária, pois, se há uma possibilidade de automação do discurso informativo, essa se dá apenas no contexto da notícia, isto é, na produção da informação primária sobre evento concreto e objetivo. Já a reportagem é resultado de operação analítica ou crítica da realidade, o que exige alto grau de subjetividade, algo, portanto, pouco propício à automação.

Um aspecto importante da notícia é que, embora se refira a acontecimentos,

não se trata de narrá-los, mas de expô-los. Em um jornal, os eventos são ordenados não por sua seqüência temporal, mas pelo interesse ou importância decrescentes. Está aí um ligeiro grau de subjetividade do qual não se pode escapar. A avaliação do que é relevante ou interessante cabe ao jornalista, levando-se em consideração, entretanto, a suposta perspectiva de quem vai ler o produto. O que não impede que esse processo se torne objetivo, já que diferentes veículos de comunicação têm seus públicos bem definidos. O jornal é pensado, portanto, para a média de seu público.

Há também que se considerar a noção intuitiva da notícia. Mesmo uma criança sabe noticiar sem recorrer, para isso, ao relato temporal. Ao chegar em casa, é mais provável que ela diga: "*meu coleguinha quebrou o braço brincando de pega-pega na hora do recreio*" em vez de fazer o seguinte relato:

"*hoje aconteceu algo diferente na escola. Quando bateu o terceiro sinal, guardamos os cadernos, pegamos a merendeira e fomos para o recreio. A Maria queria brincar de rouba-bandeira, a Anita queria jogar damas, mas decidimos brincar de pega-pega. Tiramos dois ou um para escolher quem seria o pegador. O sorteado foi Joãozinho. Aí, estávamos correndo quando Paulinho caiu. De início a gente riu, mas, como ele começou a chorar, ficamos preocupados. Joana foi chamar a professora Cida, que chegou e começou a olhar se o Paulinho tinha se machucado. Foi aí que ela descobriu que ele tinha quebrado o braço*".

Tal como num relato oral, a estrutura da notícia é lógica. Não narra nem argumenta, mas é expositiva e axiomática. LAGE (1987) divide o processo de sua produção em três fases:

- **Seleção de eventos**: no exemplo acima, ao contar à mãe sobre o acidente na escola, a criança suprimiu detalhes como a pendenga em torno de qual brincadeira escolher, ou o sorteio do pegador. Mas se não mencionasse que o menino quebrou o braço brincando de pega-pega na hora do recreio, provavelmente a mãe teria perguntado *como* e *quando;*
- **Ordenação dos eventos**: a atenção do interlocutor se fixa a partir do evento mais importante ou interessante. A importância dos outros eventos vai depender da motivação do principal. Serão, portanto, circunstanciais e explicativos;
- **Nomeação**: os nomes que se atribuem às coisas têm compromissos com os interlocutores e refletem uma escolha cultural. Dificilmente uma criança diria "*meu colega quebrou o cúbito*", assim

como um adolescente preferiria dizer "*na hora do intervalo*" para marcar o rompimento com a palavra *recreio*, que lembra infância;

1.2 – Restrições de linguagem

Somente essa estrutura objetiva já seria um facilitador para um processo de automação da notícia. Mas há ainda outra característica que aumenta as possibilidades não só da automação, mas do tratamento da notícia, inserindo-a em um Sistema de Recuperação da Informação (SRI): o caráter pragmático da linguagem jornalística. Uma vez que o Jornalismo se propõe a processar informação em escala industrial para consumo imediato, busca-se um texto não apenas fácil e rápido de ser produzido, mas com alto grau de comunicabilidade, ou seja, capaz de ser lido e compreendido pelo maior número de pessoas possível.

> "*As circunstâncias da relação entre o jornalista e o público – a pragmática dessa relação – determinam restrições específicas no código lingüístico*", afirma LAGE (1987).

As limitações do código lingüístico acontecem reduzindo-se o número de itens léxicos (palavras e expressões) e de operadores (regras gramaticais) usados na produção da notícia. A observação da prática em um jornal diário mostra que, entre as restrições que se aplicam à linguagem no texto das notícias, destacam-se:

- o uso da terceira pessoal verbal e de verbos no pretérito perfeito, futuro e presente pelo futuro do Indicativo;
- o veto de palavras e expressões inaceitáveis no registro coloquial, salvo os termos técnicos indispensáveis; por outro lado, o veto a palavras e expressões inaceitáveis no registro formal;
- a eliminação de qualificativos com sentido testemunhal e valorativo, e sua substituição, quando possível, por aferições objetivas, verificáveis;
- a construção de períodos menores e mecanismos sintáticos menos complicados;
- o recurso a fórmulas ou modelos estruturais tais como "x pessoas morreram quando...", "x milhões de reais foram perdidos porque..." etc.

2 – A estrutura lógica do *lead*

2.1 O *lead* clássico

A diferença básica entre o relato oral e uma notícia de jornal, é que, no

segundo caso, não há controle das circunstâncias em que a informação será consumida. Enquanto na comunicação oral direta o *feedback* é imediato (pelas expressões faciais e comportamento do interlocutor é possível saber se a notícia está agradando, se está sendo compreendida etc), o mesmo não ocorre no caso de uma notícia publicada por um meio de comunicação. Tal problema não é exclusivo do Jornalismo. Em verdade, desde que se inventou a escrita há a preocupação de como controlar as condições de fruição de uma mensagem.

O Jornalismo apóia-se na noção de proposição completa aristotélica (consiste do sujeito, do que lhe é predicado e das circunstâncias de predicação) para tentar minimizar essa questão. Daí se originou a síntese de Laswell que deu origem ao *lead* clássico com as perguntas *quem* fez *o que*, *a quem, quando, onde, como, por que* e *para quê*. Segundo LAGE (2004):

> "*Lead é a abertura de uma notícia: proposição completa, constituída de sujeito, verbo, complementos e circunstâncias, que se inicia pela notação mais importante ou interessante e que pode apresentar-se, no nível de sua realização sintática, por um ou mais períodos no mesmo parágrafo lógico.*"

Em sua forma clássica, contém (LAGE, 1987):

- **o sujeito**, um sintagma nominal (SN1) que pode conter um substantivo, acompanhado ou não de um artigo, adjetivo, locução adjetiva, oração adjetiva; ou ainda uma locução substantiva, uma oração integrante;
- **o predicado**, ou seja, o sintagma verbal (SV), verbo ou locução verbal, acompanhado ou não de seu complemento, um objeto direto (SN2) ou indireto (kSN3). O símbolo K representa a preposição;
- **as circunstâncias**, ou sintagmas circunstanciais (SC) de tempo, lugar, modo/instrumento, causa/conseqüência.

2.2 – Campos semânticos

Do ponto de vista técnico, a notícia não é avaliada por seu conteúdo moral, ético ou político, mas por ser o relato de um fato, de um acontecimento. E embora possamos imaginar que, de Economia a Esportes, de Moda a Política, os fatos são tão diversos quanto incomparáveis, do ponto de vista semântico podem ser resumidos em três

campos: deslocamento, transformação e enunciação,

> "*porque as coisas, na realidade, agem apenas quando se movem, transformar ou comunicam algo. Assim, podemos supor que o nível mais profundo das notícias constitui-se de sentenças cujos verbos decorrem ou se relacionam com um dos seguintes: ir, fazer, dizer*",

sustenta LAGE (2004b).

DIXON *apud* LAGE (2004b) sustenta que verbos portugueses se dispõem em campos semânticos ou grupos de itens léxicos relacionados uns com os outros pelo significado. Em cada grupo, o verbo com sentido mais geral é seguido por outros, que podem ser considerados derivações do primeiro, com a adição de traços semânticos ou variações modais. Assim é que o vasto campo semântico de *ir* inclui verbos como *partir, embarcar, levar, correr, andar, contornar, navegar, voar, pousar, aproximar-se, chegar* etc.

Os verbos do campo semântico de *fazer* referem-se ao tipo de transformação que se processa: *erguer, demolir, moldar, forjar, invadir, matar/morrer, compactar, recolher, detonar* etc. Os verbos do campo semântico de *dizer* descrevem os eventos tendo em vista a perspectiva do observador, a natureza da informação ou o tipo de codificação, como *dizer, transmitir, afirmar, negar, mandar, acrescentar, conclamar,escrever* etc.

Sendo assim, um esforço para classificar o maior número possível de verbos da língua portuguesa (as classes deveriam ser consistentes do ponto de vista semântico e oferecer informação sintática, isto é, especificar a natureza dos complementos verbais) poderia dar origem à modelagem dos verbos em padrões analógicos que pudessem ser representados por equações ou em algoritmos lógicos, ou seja, que pudessem ser computáveis. Uma proposta assim já foi feita por LAGE (2004b), que dividiu os verbos em seis classes: existência, ligação, relação, ação objetiva (mais presentes na notícia), controle e ação subjetiva.

2.3 – Como falar em automação

Para LAGE (2004), antes de ser uma proposição aristotélica completa, o *lead* é um conjunto de partes menores dotadas de sentido:

> "*Consideraremos toda sentença-lead como constituída de uma locução verbal que se reporta ao fato e em torno do qual*

se organizam locuções nominais designando actantes (personagens ou seres que atuam) do fato. Circunstâncias de tempo, lugar, modo, instrumento e finalidade são consideradas novas atribuições verbais relacionadas à totalidade da ação descrita."

Aí reside, na avaliação do autor, a fresta para a automação. Ele propõe que a sentença nuclear do lead pode ser analisada como função, no sentido matemático do termo proposto por FREGE (1978). O *lead* teria variáveis e constantes predicadas (sempre maiúsculas) e variáveis e constantes individuais (sempre minúsculas), que corresponderiam aos argumentos da função.

Exemplo:

"*Uma bomba destruiu o Palácio da Liberdade*"

a) Pode-se propor que "uma bomba" é um argumento x e "destruiu o Palácio da Liberdade" é a função E, o que resultaria: E (x)

b) Pode-se propor que "uma bomba" e "Palácio da Liberdade" são dois argumentos, enquanto "destruiu" seria a função E, levando a: E (x,m). Essa é melhor adequada ao *lead*

A frase poderia ser ampliada, como "*Uma bomba destruiu o Palácio da Liberdade ferindo centenas de pessoas*". A função continuaria sendo o verbo "destruir" e o complemento seria outro argumento. A natureza da relação entre um argumento e outro é constante, expressa, no caso, pelo verbo.

A fórmula geral de uma sentença como o *lead* é, portanto:

F (x, y, z....)

Em que F corresponde ao verbo, e x, y, z ..., aos argumentos, isto é, ao sujeito x (argumento externo) e complementos y, z.... (argumentos internos) do verbo.

3 – O tratamento da informação jornalística

O primeiro benefício da automação do discurso informativo tem uma vantagem significativa para os jornais diários: a redução do tempo de produção. Especialmente para os veículos que além da tiragem impressa também mantém sites pretensamente atualizados em tempo real, um instrumento que

acelere a produção de notícias será sempre bem-vindo. Considere-se que parte do trabalho ainda será dependente do esforço humano, já que a apuração das notícias e o levantamento dos fatos (variáveis) que possam a vir a ser usados para abastecer um sistema de produção automática de um discurso continuarão sendo feitos pelo repórter. Logo, seria falaciosa qualquer afirmação de que a automação substituiria o trabalho do jornalista.

Mas há ainda outro benefício que merece ser considerado: a automação do *lead* abre uma possibilidade para a automação da análise de assunto, uma das etapas mais cruciais do tratamento da informação. A função do analista é proceder a uma leitura técnica do documento (no caso, das notícias), determinar sua tematicidade, identificar conceitos-chave, traduzir esses conceitos para um Sistema de Recuperação da Informação de forma que, quando for necessário, o usuário possa recuperar aquela informação com a maior precisão possível.

Recuperar informações para a produção de reportagens é um dos grandes problemas do jornalismo atual. Isso porque, com exceção das "pautas frias", como são chamadas as matérias programadas com antecedência (para as quais pode-se demandar bastante tempo procurando material de referência), quando ocorre um fato inesperado que demanda a produção de reportagens especiais e matérias do tipo "relembre o caso" ou "eventos similares anteriores", tal produção deve ser feita em questão de horas, às vezes, de minutos.

Por exemplo, em meados de agosto, todos os jornais já deverão ter prontas suas matérias para o "11 de Setembro", aniversário do ataque ao Worl Trade Center, evento reconhecido como um marco da história contemporânea. Mas se um incêndio criminoso destruísse a Rocinha, maior favela da América Latina, num sábado às 17h (nesse dia, por causa da edição especial de domingo, quase todos os jornais fecham as edições às 20h) teriam os jornais o que oferecer aos seus consumidores algo além da simples notícia (provavelmente já explorada, com mais emoção e dramaticidade, em função das imagens e sons, pela mídia eletrônica)?

Nesse caso, para conseguir produzir reportagens diferenciadas, os jornais deveriam contar com um eficiente Sistema de Recuperação da Informação,

que permitisse o rápido acesso e eficiente acesso ao material arquivado, não apenas com grande revocação[3] mas com a maior precisão possível. A questão é que a maioria das empresas jornalísticas não consegue tratar seu próprio produto na velocidade que é produzido, o que demandaria equipes de bibliotecários quase tão grandes quanto a de jornalistas. Os repórteres acabam tendo que se valer das informações arquivadas em estado bruto em bancos de dados o que, como veremos a seguir, tem baixa efetividade.

3.1 – Base de dados X Sistemas de Recuperação da Informação (SRI)

Sistemas de Recuperação da Informação (SRI) são aqueles que objetivam a realização dos processos de comunicação (ARAÚJO, 1994). Visam a dar acesso às informações nele registradas e foram concebidos, por volta da década de 40, numa tentativa de solucionar o problema da "explosão informacional", como costuma ser caracterizado o crescimento exponencial de documentos e informações produzidas. Mais do que "armazenar", o objetivo do SRI's é garantir que eles sejam acessados e forneçam as respostas esperadas.

Um SRI não é um simples banco de dados, assim como uma biblioteca não é um espaço com um amontoado de livros. Para que uma biblioteca cumpra sua função de permitir o acesso à literatura pretendida pelo usuário, ela precisa ser organizada. Cada obra precisa ser tratada – analisada, classificada, catalogada – de forma que possa ser encontrada pelo usuário. Da mesma forma, as bases de dados precisam sofrer algum tipo de organização (tratamento), para que possa fornecer a informação requisitada pelo usuário a tempo e a hora.

Há várias maneiras de se subdividir um SRI, em termos de subsistemas, de funções, de processos. ARAÚJO (1994) propõe a divisão em subsistemas de entrada (seleção/aquisição, descrição, representação, organização de arquivos, armazenamento); subsistemas de saída (análise e negociação de questões,

[3] Revocação e precisão são termos dos sistemas de recuperação da informação. A revocação é a resposta a uma busca feita em um banco de dados. Se desejamos fazer uma matéria sobre a cultura na Favela da Rocinha e digitamos, em um mecanismo de busca na internet, "Favela da Rocinha", teremos uma alta revocação, com o fornecimento de milhares de páginas. A precisão, entretanto, será baixa, pois teremos todo tipo de assuntos, de violência a cidadania, de urbanismo a saúde, eventualmente passando por cultura.

estratégia de busca/recuperação, disseminação/acesso ao documento) e o subsistema de avaliação, que não se refere especificamente nem à entrada e nem à saída das informações, mas, sim, a todo o sistema. O subsistema de entrada é o mais crucial. Isso porque, se a informação não for propriamente tratada no momento de ingressar no sistema, poderá haver incorreções no subsistema de saída, isto é, quando o usuário procurar a informação, poderá não encontrá-la.

Embora seja resultado de várias operações, para este trabalho, importa destacar duas etapas do subsistema de entrada: a análise de assunto e a tradução dos conceitos encontrados para uma linguagem especializada, ou seja, a indexação. É esse processo que vai permitir a posterior recuperação da informação. Segundo ARAÚJO (1994),

> "*a base do processo de indexação é uma desconstrução do texto para uma pretensa reconstrução do conteúdo. Tal 'reconstrução' baseia-se na hipótese de um determinado número de palavras-chave (variando de 5 a 30) ser capaz de retratar o conteúdo de um documento que, nas áreas científicas têm, em média, entre 5.000 e 8.000 palavras.*"

3.2 – Linguagem natural X linguagens especiais

Por que tratar a informação jornalística? Se o fazer jornalístico privilegia a agilidade, a busca direta em um banco de dados não soluciona as questões de necessidade de informação do jornalista com mais rapidez e simplicidade? E se o problema é como selecionar, entre documentos encontrados, os mais pertinentes, em se tratando de um universo restrito (o de um jornal), e em se tratando da produção do mesmo tipo de documento (a notícia), não estaria o jornalista apto a fazer ele mesmo a seleção, já que dominaria o vocabulário específico?

A resposta é não. Porque ainda que trabalhem em editorias específicas, como Política, Economia, Cidades, Turismo, os jornalistas não são especialistas, e sim, generalistas. E também não acompanham o mesmo assunto cotidianamente. Em um mesmo dia, um repórter de Cidades (onde a variedade de assuntos a serem cobertos é maior) pode cobrir três assuntos distintos, como a greve de professores, a falta de medicamentos em determinado hospital e a liberação de verbas para reconstrução de estradas. E ainda que a greve dure 45 dias, ele pode cobri-la penas no último

dia. Como saber o que se passou nos demais?

A questão do banco de dados é complexa. Algumas buscas operadas em um banco bruto até podem dar resultados efetivos, dependendo do que se procura, mas, na maioria das vezes, a resposta não será adequada. Tome-se como exemplo uma recente invasão de sem-terra, supondo-se que, além de fazer a simples notícia, o jornal decida fazer um material analítico, com um balanço da situação agrária no País. Uma busca booleana em um banco de dados com expressões como "invasão de terras", "invasões de fazendas", "conflito agrário", ou "os sem-terra invadiram..." tem altas chances de dar bons resultados (se não se estiver esperando um aspecto específico, como escolarização de crianças sem-terra ou a produtividade das fazendas pós-invasão etc).

Mas se, depois de um parto acidental de uma criança dentro de um ônibus, o repórter quiser produzir uma matéria sobre partos inusitados, como deveria proceder à busca? Há alguma expressão-chave a que todos os repórteres que um dia tivessem feito uma matéria do gênero teriam recorrido? A procura por "parto" teria uma revocação altíssima, com precisão quase nula. Entrariam na resposta matérias não apenas de saúde, mas de Política, Esportes ou Cultura: "*aprovar essa emenda foi um parto – disse o senador Ramez Tebet*", "*querem me dispensar, mas não parto daqui sem meus salários atrasados – declarou Romário*", "*o parto de Carla Peres foi transmitido ao vivo*".

Um repórter mais sensacionalista poderia ter escrito "*um parto fora do normal*", outro, mais conservador, poderia ter escrito "*um parto inusitado*". O jornalista deveria tentar "advinhar" uma expressão utilizada em matérias a serem recuperadas. Também poderia fazer infidáveis buscas, como "*parto no elevador*", "*parto na viatura*", ou "*nasceu no elevador*", "*nasceu na viatura*", ou "*dar à luz na viatura*", "*dar à luz numa viatura*", "*dar à luz em uma viatura*" etc. E repetir as operações para cinema, parque, cachoeira, viaduto, banheiro, shopping center, escada rolante e sabe-se mais onde poderia ter havido um parto inusitado (ou não seria inusitado).

Tal confusão se dá porque o repórter estaria operando com a linguagem natural, aquela usada no cotidiano, que não tem preocupação com

a recuperação da informação. Segundo GUINCHAT e MENOU (1994),

> "*as linguagens naturais são adaptadas a formas de comunicação oral ou escrita, nas quais se estabelece entre os interlocutores uma forma de diálogo. O tempo e o espaço têm um papel importante neste diálogo. A linguagem natural pressupõe nuances, associações de idéias, expressão de emoções e de valores*".

Daí a necessidade de se "traduzir" o material produzido para uma linguagem técnica, documentária, que, segundo NOVELLINO (1996), é

> "*um instrumento de padronização da indexação, a qual visa garantir (sic) que indexadores de um mesmo sistema ou sistemas afins usem os mesmo conceitos para representar documentos semelhantes. Ela é também um instrumento de comunicação ao permitir que indexadores e usuários partilhem um mesmo vocabulário.*"

4 – A indexação via o *lead* automatizado

Respondida a questão "*por que indexar a produção jornalística*", passa-se à outra: como proceder à indexação, considerando-se o volume de matérias produzidas diariamente e a necessidade de que o material já esteja tratado no dia seguinte, para responder à agilidade do jornalismo? Como já dito anteriormente, esse esforço demandaria uma equipe de bibliotecários quase tão grande quando a equipe de jornalistas. Situação ideal, porém impraticável.

O grande volume de material a ser indexado é uma das questões com as quais a Ciência da Informação tem-se deparado. Afinal, uma das principais etapas do tratamento da informação é a análise de assunto, um esforço intelectual por natureza. Mas com o volume crescente de documentos, agora em diversos novos formatos (digitais), mesmo com o emprego da leitura técnica (estratégia de leitura que permite ao analista estabelecer qual é o assunto do documento sem ter que lê-lo integralmente, atendo-se ao resumo, título e outras propriedades do texto) torna-se necessário pensar em outro tipo de processamento que não o puramente manual.

O que se buscam são alternativas que auxiliem o profissional da informação a exercer seu ofício. Uma destas formas seria a ajuda do computador na

determinação do assunto, ou seja, a leitura automática. Segundo Pêcheux *apud* LUCAS (2000), a leitura interpretativa tem sofrido influências das linguagens lógicas, que tentam realizar o sonho da inteligência artificial, que "*buscam na semântica universal a desambigüização dos enunciados, das palavras, idealizando uma linguagem homogeneizada, legível e interpretável pelas máquinas*."

A indexação automática, ou indexação derivativa, isto é, aquela executada integralmente pelo computador mediante um algoritmo, não é, ainda, uma realidade. Mas alguns avanços têm sido feitos, como na área da lingüística, onde alguns estudos se concentraram nas teorias léxicas de enumeração de sentido. Teóricos dessa área tratam as palavras como um conjunto estático de sentidos de palavra, com marcações que ajudariam a dar informações sobre a mesma, como categoria, semântica etc. Isso possibilitaria que a leitura automática diferenciasse sentidos para a palavra raiz, por exemplo, que pode ser raiz histórica, vegetal, raiz de problemas, raiz de cabelo, raiz de uma árvore computacional binária, entre outros.

Considerando que o *lead*, além de ser o primeiro parágrafo de uma matéria jornalística é, em verdade, seu resumo, é razoável que se considere ser possível identificar o assunto de toda a matéria pela sua abertura. Além disso, como já exposto, para aumentar a comunicabilidade e facilitar a produção da mensagem, o *lead* é sujeito a uma série de restrições pragmáticas, sobretudo no uso de vocabulário e gramática. Os campos semânticos verbais também são restritos, tornando o conjunto de vocábulos a serem usados no *lead* relativamente pequeno. Segundo LANCASTER (1993), um vocabulário mais coeso, enxuto e menos específico torna mais simples o processo de análise de assunto.

Se a partir de uma série de variáveis é possível produzir automaticamente o *lead* das notícias, pode-se inferir que, partindo desse processo, também é possível acelerar o processo de indexação, promovendo uma primeira análise de assunto pela leitura automática. Pode-se esperar que ao mesmo tempo que origine uma produção sintática do enunciado final – o *lead* – também se origine daí outra produção sintática – a indexação.

Se existe uma possibilidade de automação do discurso informativo essa se dá apenas em relação à notícia, ou seja, em relação à produção de informação a partir de fato concreto e objetivo. Também na automação da indexação do produto jornalístico a automação não pode ser pensada como algo finalizador. Da mesma forma como não é possível que um programa de computador crie o *lead* de uma reportagem (pois supõe construção crítica e analítica), da mesma forma, a indexação de matérias com maior grau de subjetividade dependeria de uma melhor análise intelectual do profissional da informação. Mas, acredita-se, a primeira leitura automatizada poderia abreviar e valorizar esse processo.

Conclusão

A produção automática de discursos é uma preocupação de diferentes campos disciplinares. O que têm o Jornalismo e a Ciência da Informação a dizer neste momento?

Possibilidade apresentada mostra que a automação do *lead* clássico de notícias jornalísticas pode ser resposta não apenas para a necessidade de se acelerar o processo de produção de notícias (os jornais – e os jornalistas – estão sempre tentando superar o tempo, característica que se acentuou com a criação de *sites* pretensamente atualizados em tempo real), mas também possibilita o melhor tratamento das matérias produzidas pelos próprios jornais, com vistas à uma recuperação mais eficiente da informação.

Não sendo preocupação atual das empresas jornalísticas (melhor dizendo, sendo uma prioridade que se deixa de lado diante da dificuldade de se tratar, literalmente da noite para o dia, uma imensa produção diária de notícias), a ausência de um adequado tratamento da informação cria dificuldades para os jornalistas, que não conseguem acessar bancos de dados de matérias produzidas anteriormente e, com isso, sofrem um grande desgaste toda vez que precisam recorrer a produções antigas para produzir novo material.

Recomenda-se estudos complementares, inclusive práticos, para analisar-se melhor as possibilidades aqui levantadas. Mas ressalta-se que qualquer possibilidade de automação, tanto na produção das notícias, quando da sua análise de assunto, deve ser vista apenas

como um auxílio ao trabalho do jornalista e do profissional da informação. A beleza, e o diferencial, das duas atividades continua sendo o esforço intelectual.

REFERÊNCIAS

ARAÚJO, V.M.R.H. de. **Sistemas de recuperação da informação**. Rio de Janeiro: UFRJ, 1994.

BARDIN, L. **Análise de conteúdo**. Lisboa: Edições70, 1995. p. 27-46: Definição e relações com as outras ciências

DAHLBERG, Ingetraut. Teoria do Conceito. **Ciência da Informação**, Rio de Janeiro, v.7, n.2, p. 101-107, jul./dez. 1978.

DIAS, E.W. **Análise de assunto: percepção do usuário quanto ao conteúdo de documentos**. Texto submetido à publicação em periódico.

FUJITA, M.S.L. A identificação de conceitos no processo de análise de assunto para indexação. **Revista Digital de Biblioteconomia e Ciência da Informação**, Belo Horizonte, v.1, n.1, p. 60-90, jul./dez.2003.

GUINCHAT, C., MENOU, M. **Introdução geral às ciências e técnicas da informação e documentação**. Brasília: IBICT, 1994.

LAGE, Nilson. **Estrutura da notícia.** São Paulo: Editora Ática,1987.

_____. **Linguagem jornalística**. São Paulo: Editora Ática,1990.

_____.**O lead clássico como base para automação do discurso informativo.** Curso de Jornalismo da Universidade Federal de Santa Catarina. Disponível em <www.jornalismo.ufsc.br/bancodedados/lage-olead.html> . Acesso em 20 jun. 2004.

_____. **Modelos computáveis para alguns verbos portugueses.** Curso de Jornalismo da Universidade Federal de Santa Catarina. Disponível em <www.jornalismo.ufsc.br/bancodedados/lage-modelosverbos.html> . Acesso em 20 jun. 2004.

LARA, M.L.G. de. **A representação documentária:** em jogo a significação. São Paulo, 1993. (Dissertação, Mestrado em biblioteconomia e documentação). P. 45-61: O processo de análise e síntese de textos.

LASWELL ????? RA, M.L.G. de. **A representação documentária:** em jogo a significação. São Paulo, 1993. (Dissertação, Mestrado em biblioteconomia e documentação). P. 45-61: O processo de análise e síntese de textos.

NAVES, Madalena M. L. Estudo de fatores interferentes no processo de análise de assunto. **Perspectivas em Ciência da Informação**, Belo Horizonte, v.6, n.2, p.189-2003, jul./dez.,2001.

NOVELLINO, Maia Salet Ferreira. Instrumentos e metodologias de representação da informação. **Informação & Informação**, Londrina, v.1, n.2, p.37-45, jul./dez. 1996.