Recuperação de Informação

OLINDA NOGUEIRA PAES CARDOSO[1]

[1]UFLA – Universidade Federal de Lavras
DCC – Departamento de Ciência da Computação
Cx. Postal 37 – CEP 37.200-000 Lavras (MG)
olinda@comp.ufla.br

Resumo: Recuperação de Informação é uma área da Ciência da Computação que lida com armazenamento automático e recuperação de documentos, que são de grande importância devido ao uso universal da linguagem para comunicação. Este artigo apresenta uma visão geral dos modelos, componentes e um método de avaliação dos sistemas de recuperação de informação. São descritos os componentes de um sistema, um método de avaliação e os modelos clássicos de recuperação de informação. É apresentada a realimentação de relevantes, uma importante técnica para aumentar o desempenho dos sistemas de informações. Tópicos relacionados à área de recuperação de informação são brevemente descritos.

Palavras-chave: Modelos de recuperação de informação, bibliotecas digitais, bancos de dados textuais, realimentação de relevantes.

1 Introdução

Com o crescimento do volume de publicações, ao longo dos anos, foram desenvolvidas técnicas de recuperação de informação para responder às necessidades dos usuários de bibliotecas, tradicionais ou digitais. A ferramenta mais importante para auxiliar o processo de recuperação é denominada índice, que é uma coleção de termos que indicam o local onde a informação desejada pode ser localizada [Frakes (1992)]. Estes termos devem ser organizados de forma a facilitar sua busca.

Atualmente já não se pode falar em crescimento do volume de publicações mas em uma verdadeira explosão. As bibliotecas digitais, que são publicações armazenadas e manipuladas eletronicamente, aparecem como um paradigma para melhorar a busca e apresentação de informações desejadas. Neste contexto são estudadas técnicas de digitalização de objetos originados de fontes heterogêneas, técnicas de armazenamento, processos de busca, recuperação e apresentação de forma amigável das informações. A indexação ainda é a principal ferramenta para recuperação de informação.

A crescente complexidade dos objetos armazenados e o grande volume de dados exigem processos de recuperação cada vez mais sofisticados. Diante deste quadro, recuperação de informação apresenta a cada dia, novos desafios e se configura como uma área de significância maior.

2 Sistemas de recuperação de informação

Recuperação de informação é uma subárea da ciência da computação que estuda o armazenamento e recuperação automática de documentos, que são objetos de dados, geralmente textos. Um sistema de Recuperação de Informação (SRI) pode ser estruturado conforme a Figura 1 [Gey (1992)].

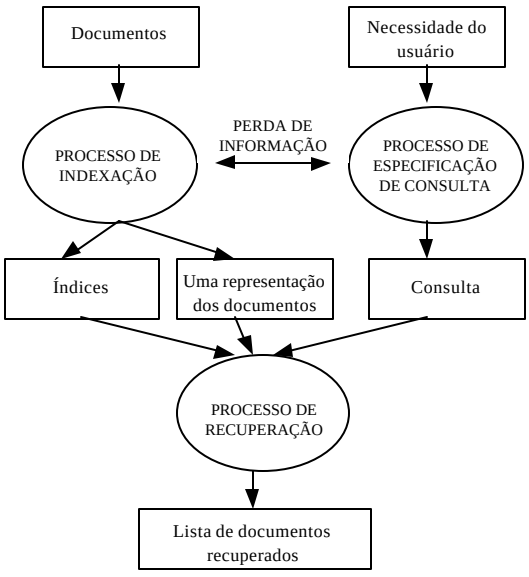

Figura 1: Componentes de um sistema de recuperação de informação

Os componentes do sistema incluem documentos, necessidades do usuário, gera a consulta formulada, e finalmente o processo de recuperação que, à partir das estruturas de dados e da consulta formulada, recupera uma lista de documentos considerados relevantes.

O processo de indexação envolve a criação de estruturas de dados associados à parte textual dos documentos, por exemplo, as estruturas de arranjos de sufixos (*PAT arrays*) e arquivos invertidos, discutidas em [Frakes (1992)]. Estas estruturas podem conter dados sobre características dos termos na coleção de documentos, tais como a freqüência de cada termo em um documento.

O processo de especificação da consulta geralmente é uma tarefa difícil. Há freqüentemente uma distância semântica entre a real necessidade do usuário e o que ele expressa na consulta formulada. Essa distância é gerada pelo limitado conhecimento do usuário sobre o universo de pesquisa e pelo formalismo da linguagem de consulta.

O processo de recuperação consiste na geração de uma lista de documentos recuperados para responder a consulta formulada pelo usuário. Os índices construídos para uma coleção de documentos e são usados para acelerar esta tarefa. Além disso, a lista de documentos recuperados é classificada em ordem decrescente de um grau de similaridade entre o documento e a consulta.

2.1 Avaliação de sistemas de recuperação de informação

Os sistemas de recuperação de informação podem ser avaliados através de consultas que fazem parte de uma coleção de referência. Um exemplo é a conhecida coleção *TIPSTER*, usada na *Text REtrieval Conference* (TREC), descrita em [Harman (1993)]. A TIPSTER é uma coleção de cerca de um milhão de documentos, obtidos de várias fontes, tais como o *Wall Street Journal*. Nesta coleção há um conjunto de consultas e para cada consulta é fornecido um conjunto ideal de documentos resposta, criado por especialistas nos temas envolvidos.

Um SRI classifica os documentos recuperados para cada consulta, de acordo com uma ordem de relevância gerando um vetor resultado. Avalia-se o SRI através da comparação das respostas geradas por este sistema e o conjunto ideal de respostas. Para isso, o vetor resultado é examinado e comparado com o conjunto ideal, obtendo-se dois índices de avaliação: precisão e revocação.

Precisão é a fração dos documentos já examinados que são relevantes, e revocação é a fração dos documentos relevantes observada dentre os documentos examinados. A avaliação do modelo de um SRI pode ser observada por um gráfico com as médias precisão x revocação. O gráfico pode ser obtido calculando-se a precisão para níveis anteriormente estabelecidos de revocação. A Figura 2 ilustra a forma geral de um gráfico precisão x revocação. Seja, N o conjunto de resposta ideal, $|N|$ o número de documentos deste conjunto e R o vetor resultado recuperado pelo SRI. Então,

$$Revocação = \frac{|N \cap R|}{|N|}$$

$$Precisão = \frac{|N \cap R|}{|R|}$$

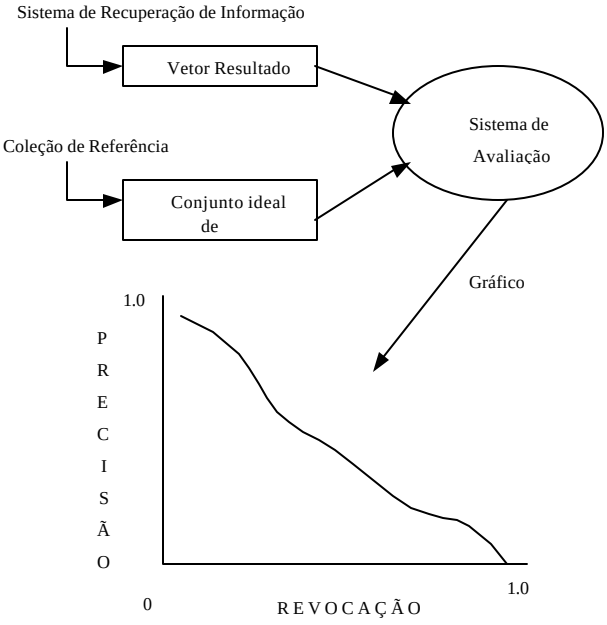

Figura 2: Processo de avaliação de um sistema de recuperação de informação

3 Modelos clássicos

Os modelos clássicos, utilizados no processo de recuperação de informação (booleano, vetorial e probabilístico) apresentam estratégias de busca de documentos relevantes para uma consulta (*query*).

Estes modelos consideram que cada documento é descrito por um conjunto de palavras chaves, chamadas termos de indexação. Associa-se a cada termo de indexação t_i em um documento d_j um peso $w_{ij} \geq 0$, que quantifica a correlação entre os termos e o documento.

Além dos modelos clássicos, modelos muito mais avançados de recuperação de informação tem sido propostos ao longo dos anos, dentre estes, destacam-se modelos baseados em bases de conhecimento [Biwas

(1987)], lógica *fuzzi* [Bookstein (1980)] e redes neurais [Kwok (1995)].

3.1 Modelo Booleano

Dada uma consulta *Q* e um conjunto de documentos considerados relevantes para a *Q*, o índice atribuído aos documentos deve indicar qual documento é mais relevante que outro, estabelecendo uma ordem de relevância. Esses índices são calculados com base na comparação entre a consulta e os documentos.

No modelo booleano os documentos recuperados são aqueles que contêm os termos que satisfazem a expressão lógica da consulta. Uma consulta é considerada como uma expressão booleana convencional formada com os conectivos lógicos *AND*, *OR* e *NOT*.

Uma maneira direta de implementar o modelo booleano seria [Salton (1989)]: assuma a existência de uma lista invertida na qual cada entrada corresponde a um termo de indexação, ademais, a entrada t_i aponta para uma lista de documentos nos quais o termo t_i ocorre. O conjunto de documentos recuperados pode ser obtido pela interseção das listas invertidas de documentos, dos termos que aparecem na consulta. Assim, somente documentos cujos termos de indexação satisfazem a consulta booleana são recuperados.

Os principais problemas do modelo booleano são a ausência de ordem na resposta, e as respostas podem ser nulas ou muito grandes. As vantagens desse modelo são a facilidade de implementação, e a expressividade completa das expressões.

3.2 Modelo vetorial

O modelo de espaço vetorial, ou simplesmente modelo vetorial, representa documentos e consultas como vetores de termos. Termos são ocorrências únicas nos documentos. Os documentos devolvidos como resultado para uma consulta são representados similarmente, ou seja, o vetor resultado para uma consulta é montado através de um cálculo de similaridade.

Aos termos das consultas e documentos são atribuídos pesos que especificam o tamanho e a direção de seu vetor de representação. Ao ângulo formado por estes vetores dá-se o nome de $\boldsymbol{q}$. O cos $\boldsymbol{q}$ determina a proximidade da ocorrência. O cálculo da similaridade é baseado neste ângulo entre os vetores que representam o documento e a consulta, através da seguinte fórmula [Salton (1988)].

$$sim(d,q) = \frac{\sum_{i=1}^{t} w_{id} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{id}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}}$$

Os pesos quantificam a relevância de cada termo para as consultas (W_{iq}) e para os documentos (W_{id}) no espaço vetorial. Para o cálculo dos pesos W_{iq} e W_{id}, utiliza-se uma técnica que faz o balanceamento entre as características do documento, utilizando o conceito de freqüência de um termo num documento. Se uma coleção possui *N* documentos e n_{ti} é a quantidade de documentos que possuem o termo t_i, então o inverso da freqüência do termo na coleção, ou *idf* (*inverse documento frequency*) é dado por:

Este valor é usado para calcular o peso, utilizando a

$$idf_i = \log \frac{N}{n_i}$$

seguinte fórmula: $W_{id} = freq(t_i,d)$ x idf_i , ou seja, é o produto da freqüência do termo no documento pelo inverso da freqüência do termo na coleção.

As principais vantagens do modelo vetorial são a sua simplicidade, a facilidade que ele provê de se computar similaridades com eficiência e o fato de que o modelo se comporta bem com coleções genéricas.

3.3 Modelo probabilístico

O modelo probabilístico descreve documentos considerando pesos binários que representam a presença ou ausência de termos. O vetor resultado gerado pelo modelo tem como base o cálculo da probabilidade de que um documento seja relevante para uma consulta. A principal ferramenta matemática do modelo probabilístico é o teorema de Bayes [Van (1979)].

O modelo probabilístico é baseado no princípio probabilístico de ordenação (*Probability Ranking Principle*), que estabelece que este modelo pode ser usado de forma ótima. Este princípio é baseado na hipótese de que a relevância de um documento para uma determinada consulta é independente de outros documentos. O princípio é o seguinte:

> "Se a resposta de um sistema de recuperação de referência a cada requisição, é uma ordem de documentos classificada de forma decrescente pela probabilidade de relevância para o usuário que submeteu a requisição, onde as probabilidades são estimadas com a melhor precisão com base nos dados disponíveis, então a efetividade geral do sistema para o seu usuário, será a melhor que pode ser obtida com base naqueles dados".

O modelo probabilístico considera um processo iterativo de estimativas da probabilidade de relevância.

Devem ser calculados: $P(+R_q|d)$ a probabilidade de que um documento d seja relevante para uma consulta q e $P(-R_q|d)$ a probabilidade de que um documento d não seja relevante para uma consulta q.

O documento d é considerado relevante para a consulta q se $P(+R_q|d) > P(-R_q|d)$, e o vetor resultado é decidido com base num fator $W_{d|q}$, definido por:

$$W_{d|q} = \frac{P(+R_q \mid d)}{P(-R_q \mid d)}$$

Este fator minimiza a média do erro probabilístico.

Através do teorema de Bayes e estimativas de relevância baseadas nos termos da consulta, pode-se chegar a seguinte equação:

$$sim(d,q) = W_{d|q} = \sum_{i=1}^{t} x_i \times W_{qi}$$

Onde:

- $x_i \in \{0, 1\}$;
- $W_{qi} = \log r_{qi} (1-s_{qi}) / s_{qi}(1-r_{qi})$;
- r_{qi} é a probabilidade de que um termo de indexação i ocorra no documento, dado que o documento é relevante para a consulta q; e
- s_{qi} é a probabilidade de que um termo de indexação i ocorra no documento, dado que o documento não é relevante para a consulta q.

O modelo probabilístico tem como vantagem, além do bom desempenho prático, o princípio probabilístico de ordenação, que uma vez garantido, resulta em um comportamento ótimo do método. Entretanto, a desvantagem é que este comportamento depende da precisão das estimativas de probabilidade. Além disso, o método não explora a freqüência do termo no documento e ignora o problema de filtragem de informação.

4 Realimentação de Relevantes

Existem várias dificuldades para que o usuário transforme suas necessidades em uma consulta devidamente formulada. Geralmente é a má formulação da consulta que prejudica o desempenho dos sistemas. Um método de abordar este problema é considerar uma forma interativa de construção da consulta, onde o usuário formula uma consulta inicial, examina o resultado diante de suas necessidades e se necessário melhora a formulação da consulta.

A estratégia mais popular para reformular consultas é chamada de realimentação de relevantes [Robertson (1976)], cuja idéia principal é como se segue. Após a montagem do vetor resultado baseado na consulta inicial, o usuário seleciona documentos de sua preferência. O sistema então seleciona termos pertencentes aos documentos selecionados e utiliza estes termos para reformular a consulta. Este processo de reformulação pode prosseguir com mais de uma interação.

A principal vantagem do método é que após a primeira formulação o usuário interage com o sistema abstraindo-se do processo de formulação, simplesmente identificando documentos como relevantes ou não. Outra vantagem é que o método provê um processo controlado de enfatizar alguns termos e diminuir a importância de outros.

5 Tópicos especiais em recuperação de informação

Nesta seção serão brevemente tratados alguns tópicos adicionais freqüentemente utilizados na modelagem de um SRI. Estes tópicos incluem: passagens, expansão de consultas, filtragem de informação, categorização e extração de informação, e visualização.

5.1 Passagens

Em recuperação de informação geralmente o usuário necessita identificar qual parte do documento retornado atende sua necessidade de informação. Uma forma de apresentar esta informação ao usuário é dividir o documento em porções menores denominadas passagens. Trabalhos recentes sugerem que, no contexto de documentos com estrutura interna complexa, evidências a nível de passagens são importantes para os sistemas de recuperação de informações. Em alguns casos, aplicar os algoritmos de recuperação a passagens, e não a documentos completos, resulta em melhor desempenho do sistema [Callan (1996)].

A divisão dos documentos em passagens pode ser feita de três formas. A primeira, considera passagens com características de hierarquia dos documentos como sentenças, parágrafos e seções. A segunda, considera passagens baseadas nas características semânticas do conteúdo de partes do documento, neste sentido as passagens agrupam porções do texto que tratam de um determinado assunto. A terceira, considera passagens como uma seqüência contígua de palavras, esse tipo de passagem é chamada de janela e o número de palavras na seqüência define o tamanho da janela.

5.2 Expansão de consultas

Um problema fundamental em recuperação de informação é que os autores nem sempre usam as mesmas palavras que os usuários para descrever o mesmo conceito [Xu (1996)].

A importância deste problema tende a diminuir com o aumento do amanho da consulta. Entretanto, em muitas aplicações, as consultas podem possuir uma pequena quantidade de termos. Um caso extremo ocorre no contexto da *Web*, onde as consultas possuem tipicamente duas palavras.

A expansão de consultas é um caminho para solucionar estes problemas. Para expandir a consulta, pode-se usar realimentação de relevantes, mas isso requer intervenção do usuário. Uma outra idéia seria a de expandir a consulta de forma automática, ou seja, sem a intervenção do usuário.

Para expandir uma consulta é preciso buscar palavras com significados semelhantes aos termos da consulta e acrescentar tais palavras à consulta original com o objetivo de melhorar o contexto da mesma. Duas abordagens podem ser adotadas: o uso de dicionários de sinônimos e o uso de palavras que co-ocorrem com os termos das consultas em documentos da coleção. No caso de dicionários de sinônimos os resultados obtidos não são em geral muito bons. Melhorias consideráveis foram alcançadas quando considerou-se análise automática de termos que co-ocorrem em documentos da coleção.

Outro tópico importante em expansão de consulta é a quantidade de termos adicionados a consulta. Nos experimentos apresentados em [Harman (1992)] o melhor desempenho foi alcançado com adições entre 20 e 40 termos, mas claramente este número depende da coleção utilizada.

5.3 Filtragem de informação

O processo de filtragem de informação consiste em analisar um fluxo de informações que chega, comparar os documentos neste fluxo com tópicos de interesse do usuário e selecionar os documentos pertinentes [Belkin (1992)]. Tipicamente o sistema funciona como um agente inteligente que seleciona os documentos do fluxo de acordo com um perfil pré-definido do usuário, geralmente estático. O problema pode ser abordado com modelagem clássica, mas existem pequenas diferenças devido ao fato dos documentos serem dinâmicos, geralmente grandes, e das necessidades do usuário serem relativamente estáticas.

Os sistemas de filtragem de informação geralmente dão uma maior ênfase na representação das necessidades do usuário, isto é, na definição do perfil do usuário, o que difere dos sistemas de recuperação de informação. Além disso, um problema que parece alcançar maior destaque em filtragem do que em recuperação de informação é a representação de dados não textuais.

Uma comparação entre recuperação e filtragem de informação é apresentada em [Belkin (1992)]. Os autores concluem que:

> "Filtragem de informação e recuperação de informação são dois lados da mesma moeda, trabalham para ajudar pessoas a obter informações necessárias para executar suas tarefas".

5.4 Categorização e extração de informação

Categorização é o processo de classificar documentos em categorias pré-definidas. Sua maior aplicação tem sido para atribuir categorias a documentos e posteriormente utilizar estas categorias para suportar recuperação e filtragem de informação.

As categorias são definidas através de um pequeno conjunto de características e tendem a ser mais estáticas que os perfis em filtragem de informação. Sistemas de recuperação de informação apresentam baixo desempenho no contexto de categorização, principalmente devido ao vocabulário restrito que descreve as categorias e o vocabulário irrestrito dos documentos [Yang (1994)].

Extração de informação é o problema de obter a partir de documentos algumas informações específicas. Como por exemplo, obter o nome de seqüestradores e de vítimas em ataques terroristas. Neste caso, a parte do documento que não é relevante pode ser ignorada. Geralmente o problema é abordado no contexto de coleções específicas. Uma abordagem para o problema é varrer o texto, buscando palavras chaves e extrair dos contextos onde ocorrem tais palavras a informação necessária. Várias alternativas de tratamento deste problema são apresentados em [Allen (1994)].

5.5 Visualização

Mesmo com as interfaces mais avançadas, com relação a interação com o usuário, expressar uma necessidade de informação é uma tarefa difícil. Existe uma distância semântica entre a real necessidade do usuários e o que ele expressa na consulta formulada. Esta distância é provocada principalmente pelo limitado conhecimento do usuário no universo da pesquisa. Além do problema de formulação da consulta, o grande volume de dados presentes nos sistemas de recuperação de informação atuais implica que a apresentação dos resultados para o usuário também é uma tarefa difícil.

Facilitar a formulação de consulta e a apresentação dos dados são problemas estudados na área de

visualização. O objetivo é desenvolv er mecanismos para apresentar visualmente os dados ao usuário, bem como permitir que este explore os dados de forma amigável. Experiências com abordagens alternativas de visualização no contexto de recuperação de informação, com melhorias de desempenho dos sistemas, são apresentadas em [Dubin (1995)], onde é feita uma análise de interfaces desenvolvidas especificamente para visualização de documentos, e [Nowell (1996)], onde são levantadas algumas alternativas de cálculos para a similaridade entre documentos e consultas, e seus efeitos na visualização dos resultados.

6 Conclusão

Neste artigo, foi apresentada uma visão geral de modelagem em sistemas de recuperação de informação, onde foram descritos os três modelos clássicos. Além dos modelos, alguns tópicos relacionados a área de recuperação de informação foram brevemente apresentados.

O estudo da área de recuperação de informação é de grande utilidade para a comunidade de sistemas de informações em geral. De fato, com a explosão do número de documentos e usuários na *Web*, modelos para recuperação precisa de informações passaram a ser de muito maior importância.

7 Referência Bibliográfica

Allen, R. B. (Ed) *ACM Transactions on Informatios Systems – Special Isuue on Text Categorization*, Vol. 12, Nº 3,1994.

Belkin, J. N. & Croft, B. W. "Information Retrieval and Information Filtering: Two sides of the same Coin?", *Comunications of the ACM*, Vol. 35, Nº 12,1992.

Biwas, G., Bezdek, J., Marques, M. & Subramanian, V. "Knowledge-Assisted Document Retrieval: II. The Retrieval Process", *Journal of the American Society for Information Science (JASIS)*, Vol. 38, Nº 2, 1987.

Bookestein, A. "Fuzzy Re quests: An Approach to Weighted Boolean Searches", *Journal of the American Society for Information Science (JASIS)*, Vol. 31, Nº 7, 1980.

Callan, J. P. "Passage-Level Evidence in Document Retrieval", *Proceedings of the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 1996.

Dubin, D. "Document Analysis for Visualization", *Proceedings of the 18th ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 1995.

Frakes, W. B. & Baeza-Yates, R. *Information Retrieval Data Structures & Algorithms*, Prentice Hall, 1992.

Gey, F. "Models in Information Retrieval". *Folders of Tutorial Presented at the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 1992.

Harman, D. "Relevance Feedback Revisited", *Proceedings of the 15th ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 1992.

Harman, D. "Overview of the Third Text REtrieval Conference(TREC-3)", http://wwwnlpir.nist.gov/TREC/t3_proceedings.html, 1993.

Kwok, K. L. "A Network Approach to Probabilistic Information Retrieval", *ACM Transactions on Information Systems*, Vol. 13, Nº 3, 1995.

Nowell, L. T., France, R. K., Hix, D., Heath, L. S. & Fox, E. A. "Visualization Search Results: Some Alternatives to Query-Document Similarity", *Proceedings of the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 1996.

Robertson, S. E. & Spark Jones, K. "Relevance Weighting of Search Terms", *Journal of the American Society for Information Science (JASIS)*, Vol. 27, Nº 3, 1976.

Salton, G. & Bucckley, C. "Term-weighting approaches in Automatic Retrieval", *Information Processing & Management*, Vol. 24, Nº 5, 1988.

Salton, G. *Automatic Text Processing: The Transformation, Analisys, and Retrieval of Information by Computer*, Addison Wesley, 1989.

Van Rijsbergen, C. J. *Information Retrieval*, Butterworths, 2nd edition, 1979.

Xu, J. & Croft, B. W. "Query Expansion Using Local and Global Document analysis", *Proceedings of the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 1996.

Yang, Y.. & Chute, C. G. "An Example-Based Mapping Method for Text Categorization and Retrieval", *ACM Transactions on Informatios Systems – Special Isuue on Text Categorization*, Vol. 12, Nº 3, 1994.