Steyvers, M. & Griffiths, T.L. (2008). Rational Analysis as a Link between Human Memory and Information Retrieval. In N. Chater and M Oaksford (Eds.) The Probabilistic Mind: Prospects from Rational Models of Cognition. Oxford University Press. pp 327-347

Chapter 15

Rational Analysis as a Link between Human Memory and Information Retrieval

Mark Steyvers*

Department of Cognitive Sciences, University of California, 3151 Social Sciences Plaza, Irvine, CA 92697-5100, USA

Thomas L. Griffiths

Department of Psychology, University of California, Berkeley, USA

Rational Analysis as a Link between Human Memory and Information Retrieval

Rational analysis has been successful in explaining a variety of different aspects of human cognition (Anderson, 1990; Chater & Oaksford, 1999; Marr, 1982; Oaksford & Chater, 1998). The explanations provided by rational analysis have two properties: they emphasize the connection between behavior and the structure of the environment, and they focus on the abstract computational problems being solved. These properties provide the opportunity to recognize connections between human cognition and other systems that solve the same computational problems, with the potential both to provide new insights into human cognition and to allow us to develop better systems for solving those problems. In particular, we should expect to find a correspondence between human cognition and systems that are successful at solving the same computational problems in a similar environment. In this chapter, we argue that such a correspondence exists between human memory and internet search, and show that this correspondence leads to both better models of human cognition, and better methods for searching the web.

Anderson (1990) and Anderson and Schooler (1991, 2000) have shown that many findings in the memory literature related to recognition and recall of lists of words can be understood by considering the computational problem of assessing the relevance of an item in memory to environmental cues. They showed a close correspondence between memory retrieval for lists of words and statistical patterns of occurrence of words in large databases of text. Similarly, other computational models for memory (Shiffrin & Steyvers, 1997), association (Griffiths *et al.*, 2007), reasoning (Oaksford & Chater, 1994), prediction (Griffiths & Tenenbaum, 2006) and causal

induction (Anderson, 1990; Griffiths & Tenenbaum, 2005; Steyvers *et al.*, 2003) have shown how our cognitive system is remarkably well adapted to our environment.

Anderson's (1990) analysis of memory also showed for the first time that there are fundamental connections between research on memory and information retrieval systems. Because information retrieval systems and human memory often address similar computational problems, insights gained from information retrieval systems can be helpful in understanding human memory. For example, one component of Anderson's first rational memory model involved calculating the predictive probability that items will re-occur given their historical pattern of occurrences. The solution to this problem was based on information retrieval models developed for library and file systems (Burrell, 1980; Salton & McGill, 1983). Just as it is useful to know the probability that a book will be needed in order to make it available in short-term or off-site storage, it is useful to know whether a fact is likely to be needed in the future when storing it in memory.

Modern information retrieval research provides new tools for modeling the environment in which human memory operates, and new systems to which human memory can be compared. An important innovation has been the introduction of *statistical language models* to capture the statistics of the regularities that occur in natural language (e.g., Croft & Lafferty, 2003; Ponte & Croft, 1998). The goal of language modeling is to exploit these regularities in developing effective systems to assess the relevance of documents to queries. Probabilistic topic models (e.g., Blei *et al.*, 2003; Griffiths & Steyvers, 2004; Griffiths *et al.*, 2007; Hoffman, 1999; Steyvers & Griffiths, 2006; Steyvers *et al.*, 2006) are a class of statistical language models that automatically infer a set of topics from a large collection of documents. These models allow each document to be expressed as a mixture of topics, approximating the semantic themes present in those documents. Such topic models can improve information retrieval by matching queries to documents at a semantic level (Blei *et al.*, 2003; Chemudugunta *et al.*, 2007; Hoffman, 1999). Another important problem in information retrieval is dealing with the enormous volume of data available on the world wide web. For any query, there might be a very large number of relevant web pages and the task of modern search engines is to design effective algorithms for ranking the importance of webpages. A major innovation has been the PageRank algorithm, which is part of the Google search engine (Brin & Page, 1998). This algorithm ranks web pages by computing their relative importance from the links between pages.

In this chapter, we use these innovations in information retrieval as a way to explore the connections between research on human memory and information retrieval systems. We show how PageRank can be used to predict performance in a fluency task, where participants name the first word that comes to mind in response to a letter cue. We also give an example of how cognitive research can help information retrieval research by formalizing theories of knowledge and memory organization that have been proposed by cognitive psychologists. We show how a memory model that distinguishes between the representation of gist and verbatim information can not only explain some findings in the memory literature but also helps in formulating new language models to support accurate information retrieval.

A Probabilistic Approach to Information Retrieval

Search engines and human memory are both solutions to challenging retrieval problems. For a search engine, the retrieval problem is finding the set of documents that are most relevant to a user query. In human memory, the retrieval problem can be construed in terms of assessing the relevance of items stored in the mind to a memory probe (either internally generated or based on environmental cues). The common structure of these problems suggests a simple analogy between human memory and computer-based information retrieval: items stored in memory are analogous to documents available in a database of text (such as the world-wide web) and the memory probe is analogous to a user query. In this section, we explore how retrieval problems of this kind can be solved using statistical inference, following Anderson (1990).

Using notation appropriate to information retrieval, the problem is to assess $P(d_i|q)$, the probability that a document d_i is relevant given a query q. The query can be a (new) set of words produced by a user or it can be an existing document from the collection. In the latter case, the task is to find documents similar to the given document. In the context of memory retrieval, the term q corresponds to the memory probe and $P(d_i|q)$ is the conditional probability that item d_i in memory is relevant to the memory probe. Let us assume that there are D documents in the database and the goal is to retrieve some set of the most relevant documents as assessed by $P(d_i|q)$. This probability can be computed using Bayes' rule, with

$$P(d_i|q) \propto P(q|d_i)P(d_i) \tag{1}$$

where $P(d_i)$ gives the prior probability that an item will be relevant (before any query or cue is issued), and $P(q|d_i)$ is the probability of observing the query if we assume that item d_i was the item that was needed, also known as the 'likelihood.'

The prior probability, $P(d_i)$, can be used to capture the idea that not all items are equally important, with some items being more likely to be the target of retrieval. In search engines, this prior probability is often computed from the link structure between documents. For example, the PageRank algorithm assumes that if a document is linked to by many other important documents, then it is likely to be important. The importance of a document, also known as its PageRank, can be conceptualized as the prior probability of a document being relevant to any particular query. We will return to this idea in the next section when discussing the PageRank algorithm and its application to memory retrieval. In the rational memory model (Anderson, 1990; Anderson & Schooler, 1991, 2000), the prior probability of an item in memory being important was computed from its historical usage pattern, under the assumption that if items were recently accessed, they are likely to be accessed again. Anderson showed that this 'history' factor can explain the effects of spacing and repetition of items on retention.

The likelihood, $P(q|d_i)$, reflects how well a particular document matches a search query or cue. In the context of information retrieval, this can be evaluated using a *generative model* that specifies how the words in the query can be generated from a statistical language model that is derived separately for each document d_i. For example, probabilistic topic models (Blei *et al.* 2003; Griffiths & Steyvers, 2004; Griffiths *et al.*, 2007;

Hoffman, 1999; Steyvers & Griffiths, 2006; Steyvers *et al.*, 2006) assume that each document can be described by a mixture of topics where the topics are derived from an analysis of word occurrences in a large database of text – relevant documents have topic distributions that are likely to have generated the set of words associated with the query. We will return to this idea in a later section. In the rational memory model (Anderson, 1990; Anderson & Schooler, 1991, 2000), this likelihood term was referred to as the 'context' factor, where the context represented the information available at test to probe memory. This factor was evaluated using a simple generative model for the properties of items stored in memory.

Equation (1) forms part of a simple schema for solving retrieval problems: compute the posterior probability that each item is relevant, combining its prior probability of being relevant with a likelihood reflecting its relationship to the query or cue, and then return the items with highest posterior probability. This schema can be used to solve the retrieval problems faced both by internet search engines and by human memory, suggesting that it may be possible to find parallels between the two. We explore this possibility in the next two sections, focusing on the role of the prior in the first, and then turning to the likelihood in the second.

Google and the Mind: Predicting Fluency with PageRank

Many search engines produce a response to a query in two stages, first identifying the set of webpages that contain the words in the query, and then ordering those pages according to the pre-computed output of a ranking algorithm. These two stages can be mapped onto the two parts of the right hand side of (1). The first stage corresponds to an assumption that the likelihood, $P(q|d_i)$, has some constant value for any page containing the query and is zero otherwise. This guarantees that only pages containing the query will have non-zero posterior probabilities, and means that the posterior probability of each page containing the query is directly proportional to its prior probability. The second stage, ordering the pages, thus reveals the prior probability assigned to each page: if the solution to the retrieval problem is to return the pages with highest posterior probability, and the posterior probability of the candidate pages is proportional to their prior probability, then a ranking algorithm implicitly assigns a prior probability to each page.

The correspondence between ranking algorithms and priors means that the prior probability that a webpage will be relevant to a user plays a central role in internet search. This raises a simple question: how should such prior probabilities be computed? While the details of the ranking algorithms used by commercial search engines are proprietary, the basic principles behind the PageRank algorithm used in the Google search engine have been published (Brin & Page, 1998). The algorithm makes use of two key ideas: first, that links between webpages provide information about their importance (and hence their probability of being the webpage that a user might seek), and second, that the relationship between importance and linking is recursive.

In addition to carrying information about different topics, webpages contain sets of links connecting them to other pages, as shown in Fig. 15.1(a). Given an ordered set of n pages, we can summarize the links between them with a $n \times n$ matrix L, where L_{ij}

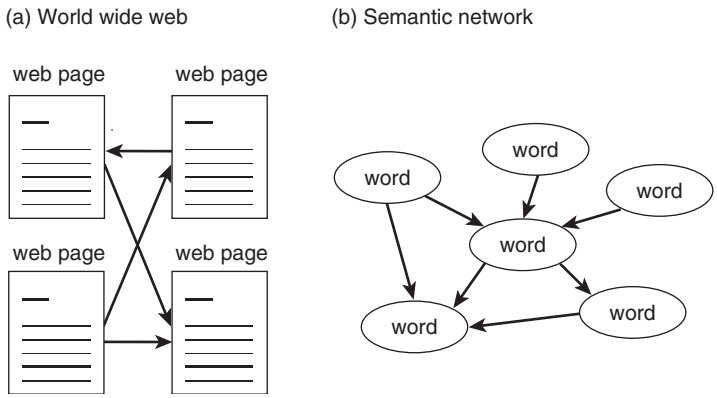

Fig. 15.1. (a) A set of webpages form a directed graph, where the nodes are pages and the edges are links. (b) Words in a semantic network also form a directed graph where the edges represent associative connections between words.

indicates that a link exists from webpage *j* to webpage *i* (the adjacency matrix of the underlying graph). This matrix provides a way to define the importance of a webpage. If we assume that links are chosen in such a way that higher importance pages receive more links, then the number of links that a webpage receives (in graph-theoretic terms, its 'in-degree') could be used as a simple index of its importance. Using the *n*-dimensional vector **p** to summarize the importance of our *n* webpages, this is the assumption that $p_i = \sum_{j=1..n} L_{ij}$.

PageRank goes beyond this simple measure of the importance of a webpage by observing that a link from a highly important webpage should be a better indicator of importance than a link from a webpage with little importance. Under such a view, a highly important webpage is a webpage that receives many links from other highly important webpages. We might thus imagine importance as flowing along the links of the graph shown in Fig. 15.1(a). If we assume that each webpage distributes its importance uniformly over its outgoing links, then we can express the proportion of the importance of each webpage traveling along each link using a matrix **M**, where $M_{ij} = L_{ij} / \sum_{k=1..n} L_{kj}$. The idea that highly important webpages receive links from highly important webpages implies a 0recursive definition of importance, and the notion of importance being divided uniformly over outgoing links gives the equation

$$\mathbf{p} = \mathbf{M}\mathbf{p} \tag{2}$$

which identifies **p** as the eigenvector of the matrix **M** with the greatest eigenvalue. The PageRank algorithm computes the importance of webpages by finding a vector **p** that satisfies this equation (ignoring a slight modification to take into account the possibility that a sequence of webpages forms a closed loop).

While the recursive definition of PageRank makes clear its assumptions about how linking affects importance, some intuitions about the factors influencing the PageRank of a page can be gained by considering an alternative route to the same formal result

(Brin & Page, 1998). We can define a random walk on the world wide web by assuming that a user starts at a randomly chosen web page, and then keeps clicking on links chosen uniformly at random from the set of links on the page reached after every click. This random walk is a Markov chain, and standard results in the mathematical theory of Markov chains indicate that, in the long run, the probability that this user lands on a particular webpage will be proportional to its PageRank.

Applying PageRank to Semantic Networks

The idea that that the pieces of information that are the targets of retrieval are connected to one another is not exclusive to web pages – it also appears in cognitive psychology. In an associative semantic network, such as that shown in Fig. 15.1(b), a set of words or concepts are represented as nodes connected by edges that indicate pairwise associations (e.g., Collins & Loftus, 1975). If we take this to be the representation of the knowledge on which retrieval processes operate, human memory and search engines thus address a similar computational problem: identifying the items relevant to a query from a large network of interconnected pieces of information. The empirical success of the Google search engine indicates that PageRank constitutes an effective solution to this problem. This raises the tantalizing possibility that the link structure of semantic networks might provide a guide to the relative importance of pieces of information, or, equivalently, an estimate of the prior probability with which a particular word or concept might be needed. In particular, it suggests that by computing the PageRank of the nodes in a semantic network, we might be able to predict the prominence of the corresponding words and concepts in memory.

In order to explore the possibility of a correspondence between PageRank and human memory, we constructed a task that was designed to closely parallel the formal structure of internet search (Griffiths *et al.* in press). Specifically, we wanted a task in which people had to produce items from memory that matched some query, with the hope that in doing so their responses would reflect the prior probability assigned to each item being needed. To this end, we showed participants a letter of the alphabet (the query) and asked them to say the first word that came into their head that begins with that letter (the relevant items). In the literature on human memory, such a task is used to measure fluency – the ease with which people retrieve different facts from memory, which can useful to diagnose neuropsychological and psychiatric disorders (e.g., Lezak, 1995). Each subject in the experiment gave fluency responses for 21 letters of the alphabet (excluding low frequency letters). The results were pooled across fifty subjects and responses that were given by only a single subject were excluded. Table 1 shows a sample of responses given for the letter 'd.'

Our goal was to determine whether people's responses could be predicted by PageRank computed from a semantic network constructed from word association norms collected by Nelson *et al.* (1998). These norms were collected by asking participants to name the first word that came into their head when presented with a cue in the form of another word. The norms list the associates that people produced for 5,018 words, and were collected in such a way that each word named at least twice as an associate also appears as a cue. From these norms, we constructed a directed graph

in which each word was represented as a node, and an edge was introduced from each word to its associates. We then applied the PageRank algorithm to this graph.

In order to evaluate the performance of PageRank, we used several alternative predictors as controls. In one control, we compared the performance of PageRank to more conventional frequency-based measures, based on the Kucera–Francis (KF) word frequency (Kucera & Francis, 1967). Word frequency is widely used as a proxy for fluency in word recognition studies (e.g., Balota & Spieler, 1999; Plaut, *et al.*, 1996; Seidenberg & McClelland, 1989; see also Adelman *et al.*, 2006) and to set the prior probability of items in rational models of memory (Anderson, 1990). Another control was a semantic network measure that was not based on a recursive definition of importance: the in-degree of each node in the semantic network. This is the frequency with which the word was named as a response in the word association norms. The in-degree of nodes in an associative semantic network has previously been used as a predictor in a number of episodic memory studies (McEvoy *et al.*, 1999; Nelson *et al.*, 2005). In-degree differs from PageRank only in the assumption that all incoming links should be given equal weight when evaluating the importance of an item, rather than being assigned weights based on the importance of the items from which they originate.

For each letter of the alphabet, we identified all words contained in the norms that began with that letter, and then ordered the words by each of the three predictors, assigning a rank of 1 to the highest-scoring word and increasing rank as the predictor decreased. A sample of the rankings for the letter 'd' produced by PageRank, KF frequency and in-degree is shown in Table 15.1. To compare performance of these three

Table 15.1. Most frequent responses in the fluency task for the letter 'd' and the rankings given by PageRank, In-degree and KF frequency.

Human responses		**PageRank**		**In-degree**		**KF Frequency**	
DOG	(19)	DOG	(19)	DOG	(19)	DO	(2)
DAD	(16)	DARK	(3)	DEATH	(1)	DOWN	(4)
DOOR	(5)	DRINK	(1)	DRINK	(1)	DAY	(2)
DOWN	(4)	DOWN	(4)	DIRTY	(0)	DEVELOPMENT	(0)
DARK	(3)	DEATH	(1)	DARK	(3)	DONE	(1)
DUMB	(3)	DOOR	(5)	DOWN	(4)	DIFFERENT	(0)
DAY	(2)	DAY	(2)	DIRT	(0)	DOOR	(5)
DEVIL	(2)	DIRTY	(0)	DEAD	(0)	DEATH	(1)
DINOSAUR	(2)	DIRTY	(0)	DANCE	(0)	DEPARTMENT	(0)
DO	(2)	DEAD	(0)	DANGER	(1)	DARK	(3)

Note: The numbers between parentheses are frequencies in human responses. All responses are restricted to the words in the word association norms by Nelson *et al*. (1998).

predictors, we compared the median ranks. The median rank assigned by PageRank was 13, as compared to 17 for in-degree and 43 for word frequency, reflecting a statistically significant improvement in predictive performance for PageRank over the controls.

The results of this experiment indicate that PageRank, computed from a semantic network, is a good predictor of human responses in a fluency task. These results suggest that the PageRank of a word could be used in the place of more conventional frequency-based measures when designing or modeling memory experiments, and support our argument that the shared problem faced by human memory and internet search engines might result in similar solutions. One way to explain the advantage of PageRank might be to return to the idea of random walks on a graph. As mentioned above, a random internet surfer will select webpages with probabilities proportional to their PageRank. For semantic networks, the PageRank of a word is proportional to the probability of selecting that word if participants started at a random word in the semantic network and proceeded to search their memories by following associative links until they found a word that matched the query (see Griffiths et al., in press, for details).

The fluency task focused on one important component in retrieval, the prominence of different words in human memory, as should be reflected in the prior $P(d_i)$. By using a letter matching task, for which the word response can either be true or false, we purposefully minimized the influence of the $P(q|d_i)$ likelihood term in (1). However, in more typical retrieval tasks, queries can relate in many ways to items stored in memory. In addition to the *form-based* matching that was emphasized in the letter-matching task, many retrieval tasks require *content-based* matching where the query and items in memory are matched at a conceptual level. In the next section, we consider the computational problem of assessing $P(q|d_i)$ using both form-based and content-based matching strategies.

Topic Models to extract Verbatim and Gist information

In both memory and information retrieval research, one of the main problems is to specify how relevant information can be retrieved in the context of a user query or environmental cues. Memory researchers have proposed that the memory system assesses relevance at two levels of generality: verbatim and gist (Brainerd *et al.*, 1999; Brainerd *et al.*, 2002; Mandler, 1980). The gist-level representation is based on a high-level semantic abstraction of the item to be stored, whether it is a sentence, conversation or document. This gist level information can be used to disambiguate words or retrieve semantically relevant concepts during reading (Ericsson & Kintsch, 1995; Kintsch, 1988; Potter, 1993). At the verbatim level, information is stored and retrieved relatively closely to the raw physical form in which it was received and might include the specific choice of words and physical characteristics related to font and voice information. While it is probably an oversimplification to propose that the memory system utilizes only two levels of abstraction to encode and retrieve information, the distinction between gist and verbatim information has been useful to understand, at least at a conceptual level, a variety of findings in memory and language research. However, these models leave open the question of exactly how verbatim and gist level information is encoded in memory.

In information retrieval, the relevance of a query to documents can be assessed using a variety of techniques that focus on different levels of abstraction of the information contained in the document and query. The simplest keyword matching strategies do not attempt any abstraction and focus on the exact word matches between documents and queries. A widely used keyword-matching retrieval technique is based is on the term-frequency, inverse-document-frequency (TF-IDF) method (Salton & McGill, 1983). The relevance of a document is related to the number of exact word matches and inversely weighted by the number of times the query terms appear in documents across the database. One problem of this technique is that it can be overly specific. It can give low relevance scores to documents that contain words semantically related to the query. To improve the generalization in retrieval, dimensionality-reduction techniques have been developed to extract a lower-dimensional description for documents that utilizes the statistical regularities of words in natural language. This has led to techniques such as Latent Semantic Indexing (LSI; Deerwester *et al*., 1990; Landauer & Dumais, 1997), and probabilistic analogues such as Probabilistic Latent Semantic Indexing (PLSI; Hoffman, 1999) and Latent Dirichlet Allocation (LDA; Blei *et al*., 2003; Griffiths & Steyvers, 2004). The idea is that queries and documents can be matched in the lower-dimensional space, which often leads to higher-level semantic matches. However, in come cases these dimensionality-reduction techniques lead to *over-generalization*. Because the matching of query and document takes place entirely in the lower-dimensional 'semantic' space, all details about the individual words in query and documents are lost in this comparison. It is possible, however, that some of the individual words in the query or document were essential to assess relevance.

The difficult issue of deciding on an appropriate level of generalization to assess relevance forms an important parallel between problems studied by memory and information retrieval researchers. In the context of human memory, should information in memory be relevant only when it exactly matches the environmental cues (using verbatim information) or should the retrieval process allow some generalization in the retrieval process (using gist)? Similarly, in information retrieval, should the relevance of documents to queries be assessed more on the level of exact matches (e.g., keyword matching strategies) or should there be some attempt to extract a more general representation of documents and queries to allow for conceptual level matches?

In this section, we consider the computational problem of balancing the trade-off between specificity and generality. We will start with a description of probabilistic topic models that focus on extracting only gist-based descriptions for each document using low-dimensional semantic representations. We then introduce an extension of these models, the dual-route topic model that augments these gist-based representations with document specific representations based on specific keyword occurrences in documents. We illustrate how this model can be used to explain several findings in the memory literature such as false memory and semantic isolation effects. We will also show how this model leads to improved performance in information retrieval.

Topic Models

Topic models such as PLSI and LDA are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model is a *generative model* for documents: it specifies a simple probabilistic procedure by which documents can be generated. In a standard topic model, to make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. To introduce notation, we will write $P(z|d)$ for the multinomial distribution over topics given document d, and $P(w|z = t)$ for the multinomial distribution over words w given a specific topic t. In a standard topic model, the distribution of words in document d can be decomposed as a finite mixture over T topics as follows:

$$P(w \mid d) = \sum_{t=1}^{T} P(w \mid z = t)P(z = t \mid d) \tag{3}$$

In this model, the $P(w|z = t)$ term indicates which words are important for topic t and $P(z = t|d)$ gives the importance of a particular topic in document d, which can be used as a representation of the content or gist of that document. In the LDA model, these multinomial distributions have associated priors, chosen to be Dirichlet distributions. The hyperparameters of the Dirichlet distributions indicate which kinds of multinomial distributions are likely, and control the degree of smoothing of the word counts in topics and topic counts in documents.

Given the observed words in a set of documents in a large corpus, we would like to know what set of topics is most likely to have generated the data. This involves inferring the probability distribution over words associated with each topic, $P(w|z)$, and the distribution over topics for each document, $P(z|d)$. Several statistical inference techniques have been developed to infer these distributions from large text corpora. The simulations discussed in this chapter utilized an efficient Gibbs sampling technique based on Markov chain Monte Carlo (Griffiths & Steyvers, 2004). We will not discuss the details of this procedure but we refer the interested reader to an introductory treatment by Steyvers and Griffiths (2006).

As an example of the topics that can be extracted with the topic model, we applied the topic model with $T = 1{,}500$ topics to the TASA corpus, a collection of over 37,000 text passages from educational materials (e.g., language & arts, social studies, health, sciences) collected by Touchstone Applied Science Associates (see Landauer *et al.*, 1998). Several topic-word distributions $P(w|z = t)$ are illustrated in Fig. 15.2. The figure shows the nine words that have the highest probability under each topic. The particular topics shown in the figure relate to various themes in agriculture and biology.

In the standard topic model, each document is described by a distribution over topics which represent the gist of a document but information about particular words is lost. For example, suppose we need to encode the following list (i.e., document) of words: PEAS, CARROTS, BEANS, SPINACH, LETTUCE, TOMATOES, CORN, CABBAGE, and SQUASH. If we encode this list as a distribution over 1,500 topics, only a few topics would receive high probability. For example, one possible distribution for this list

Topic 32	Topic 41	Topic 543	Topic 816	Topic 1321	Topic 1253
VEGETABLES	TOOLS	FARMERS	MEAT	NUTRIENTS	PLANTS
FRUITS	TOOL	CROPS	BEEF	ENERGY	PLANT
POTATOES	CUTTING	FARMING	EAT	FATS	LEAVES
FRUIT	HAND	FARMS	COOKED	VITAMINS	SEEDS
POTATO	CUT	FARM	PORK	CARBOHYDRATES	SOIL
TOMATOES	DRILL	LAND	MEAL	FOOD	ROOTS
FRESH	CHISEL	CROP	SAUCE	VITAMIN	FLOWERS
ORANGES	CARPENTER	AGRICULTURE	BREAD	MINERALS	WATER
ORANGE	METAL	GROW	COOKING	NEED	FOOD

Fig. 15.2. Example topic distributions extracted from the TASA corpus using a topic model with 1,500 topics. For each topic, the nine most likely words are shown in order of probability.

would be to give probability 0.77, 0.17, and 0.06 to topics 32, 543, and 1,253, respectively, and zero probability to all other topics. This encoding would capture the idea that the list of words contained semantic themes related to *vegetables* and *farming*. However, this encoding would not allow accurate reconstruction of the specific words that were presented. If we use (3) to reconstruct the list with these topic weights, words that were not presented on the list, such as VEGETABLES and POTATO might receive relatively higher probability. While it is a desirable feature of the model to generalize beyond the specific words on a list, what is needed is a model-based encoding that tempers this generalization with a representation for the specific words present on the list.

Dual Route Topic Models

We developed the *dual-route topic model* to capture both the specific and general aspects of documents. This model is an extension of the LDA model that allows words in documents to be modeled as either originating from general topics, or from a distribution over words that is specific for that document. We will refer to this distribution as the *special word* distribution. An important assumption in the model is that each word originates from a single route only, but there can be uncertainty about the route allocation. Each word token in a document has an associated random variable x, taking value $x = 0$ if the word w is generated via the topic route, and value $x = 1$ if the word is generated as a special-word route. The variable x acts as a switch. If $x = 0$, the standard topic mechanism is used to generate the word: a topic is sampled from the topic distribution associated with the document and a word is sampled from the topic. On the other hand, if $x = 1$, words are sampled from the special-word distribution specific to the document. We model this as multinomial with a symmetric Dirichlet prior. The switch variable x is sampled from a document-specific Bernoulli variable λ with a symmetric Beta prior. The random variable λ determines the proportion of words associated with the special word and topic route within a document. The model specifies the following probability distribution over words in a document:

$$P(w \mid d) = P(x = 0 \mid d)\sum_{t=1}^{T} P(w \mid z = t)P(z = t \mid d) + P(x = 1 \mid d)P'(w \mid d) \qquad (4)$$

where $P'(w|d)$ is the special word distribution associated with document d. Note that the model explains word occurrences as a mixture of two routes, the topic model route weighted by $P(x = 0|d)$ and the special word route weighted by $P(x = 1|d)$. If $P(x = 1|d) = 0$, the model is identical to the LDA model in (3). On the other hand, if $P(x = 1|d) = 1$, the model is identical to a unigram word model. By mixing these two components, the model allows a flexible balance between modeling general and specific aspects of documents. The latent variables in the model include the terms $P(z|d)$ and $P(w|z)$ associated with the topic model and new terms $P(x|d)$ and $P'(w|d)$. As with standard topic models, Gibbs sampling can be used to infer these distributions (see Chemudugunta *et al.*, 2007, for details).

Explaining Semantic Isolation Effects

The distinction between verbatim and gist level information can be useful to understand a number of findings in the memory literature, such as the semantic isolation effect. This effect is related to the classic finding by Von Restorff (1933) that information that stands out from the context is better remembered. Von Restorff effects can be based on physical or semantic characteristics, by presenting a word on a list in a unique color or font or drawing a word from a novel semantic category. Semantic isolation effects occur when words that semantically stand out from the list are better remembered.

Early explanations of the isolation effect focused on the role of attention (Jenkins, 1948) and surprise (Green, 1956). In this account, the unexpected isolated word leads to an increase in attention which enhances the encoding of the item. However, studies have shown that the isolate is not (always) rehearsed or attended more (e.g. Dunlosky *et al.*, 2000). Also, this account cannot explain the continued presence of isolate effects even when the isolate is presented as the first word in the list. In this case, no expectations about the list contents can have been built up yet when processing the first item. An alternative account focuses on the role of memory organization with the idea that the isolate is encoded in qualitatively different ways compared to the background items (Bruce & Gaines, 1976; Fabiani & Donchin, 1995). The dual route memory model allows a computational account for the semantic isolation consistent with this proposal. In the model, the memory system utilizes qualitatively different encoding resources to encode isolate and background items. The topic route stores the gist of the list and the special-words route stores specific words such as the isolate word.

To illustrate the dual-route topic approach, we applied the model to experimental data gathered by Hunt and Lamb (2001). They compared recall performance for two lists of words, illustrated in Fig. 15.3(a). The outlier lists consisted of nine words from one category (e.g., *vegetables*) and one target word (e.g., HAMMER) from another category, whereas the control list embedded the target word in a background context that is semantically consistent. As shown in Fig. 15.3(b), Hunt and Lamb found that recall for the target word is much higher in the isolate condition, illustrating the semantic isolation effect. The finding that the target item is recalled about as well as the background items in the control list shows that this isolation effect needs to be explained by the difference in context, and not by particular item characteristics (e.g., orthography or word frequency).

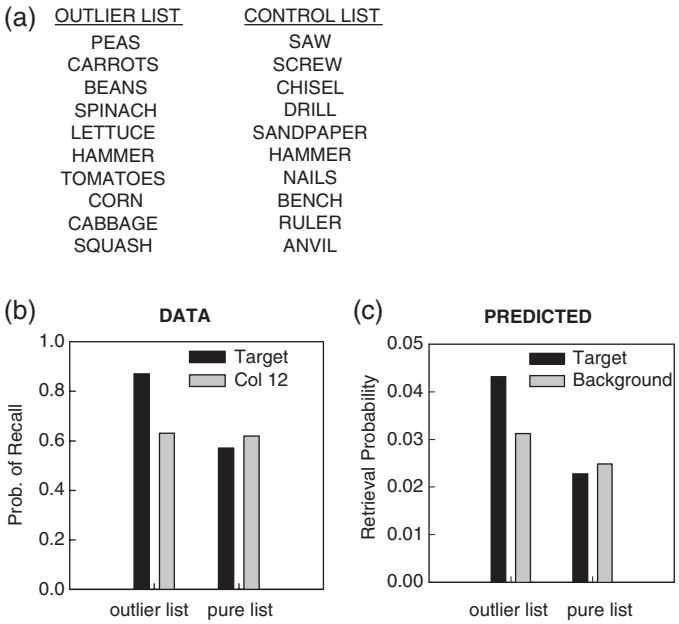

Fig. 15.3. (a) Two example lists used in semantic isolation experiments by Hunt and Lamb (2001). The outlier list has one target word (HAMMER), which is semantically isolated from the background. The control list uses the same target word in a semantically congruous background. (b) Data from Experiment 1 of Hunt and Lamb (2001) showing the semantic isolation effect (c). The predictions of the dual-route topic model.

We encoded the outlier and control lists with the dual-route topic model. To simplify the simulations, we used the same 1,500 topics illustrated in Fig. 15.2 that were derived by the standard topic model. We therefore inferred the special word distribution and topic and route weights for this list while holding fixed the 1,500 topics. We also made one change to the model. Instead of using a Dirichlet prior for the multinomial of the special-word distribution that has a single hyperparameter for all words, we used a prior with hyperparameter values that were higher for words that are present on the list than for words that were absent (0.001 and 0.0001, respectively). This change forces the model to put more a priori weight on the words that are part of the study list.

Figure 15.4 shows the model encoding for the isolate list shown in Fig. 15.3(a). The most likely topic is the vegetable topic, with smaller probability going toward the farming and tools topics, reflecting the distribution of semantic themes in the list. The special word distribution gives relatively high probability to the word HAMMER. This happens because the model encodes words either through the topic or special word route and the probability of assigning a word to a route depends on how well each route can explain the occurrence of that word in the context of other list words.

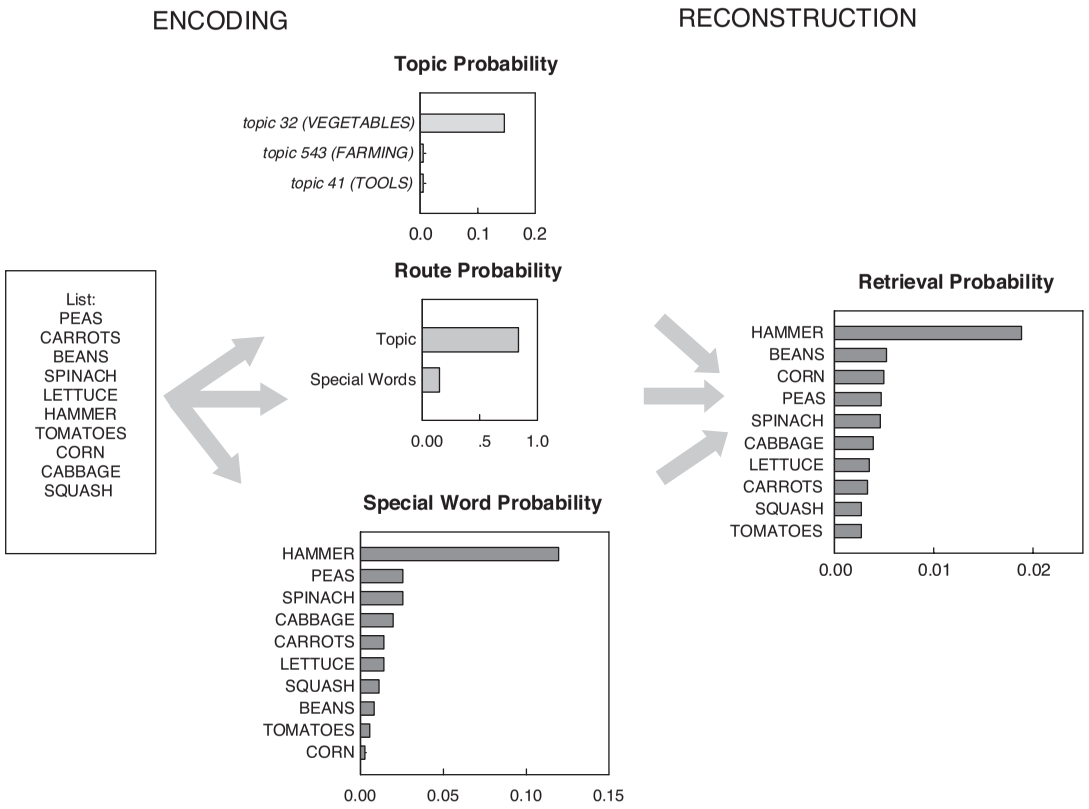

Fig. 15.4. Example encoding and reconstruction of a list of words with the dual-route topic model. Note that the topic distribution is truncated and only shows the top 3 topics. Similarly, the special-word and retrieval distributions only show the top 9 nine words from a vocabulary of 26,000+ words.

Because most of the vegetable-related words can be explained by the topic route, these words will receive lower probability from the special-word route. On the other hand, the word HAMMER, which is semantically isolated from the vegetable words cannot be explained well by the topic route, which makes it more likely to be associated with the special-word route. To simulate recall, (4) can be applied to calculate the posterior predictive probability over the whole vocabulary (26,000+ words) using the model encoding. We will refer to this as the retrieval distribution. The retrieval distribution shown in Figure 4 shows an advantage for the isolate word. This occurs because the special-word distribution concentrates probability on the isolate word, which is preserved in the reconstruction using both routes (the topic route distributes probability over all words semantically related to the list, leading to a more diffuse distribution). Figure 15.3(c) shows the model predictions for the experiment by Hunt and Lamb (2001), which exhibits the same qualitative pattern as the experimental data. Note that the retrieval probability can only be compared qualitatively to the observed recall probability. In order to fully simulate recall, we would have to implement a sampling process with a stopping rule to simulate how human participants typically produce only a subset of words from the list. For reasons of simplicity, we chose not to implement such a sampling process.

Explaining False Memory effects

The dual-route topic model can also be used to explain false memory effects (Deese, 1959; McEvoy *et al.*, 1999; Roediger *et al.*, 2001). In a typical experiment that elicits the false memory effect, participants study a list of words that are associatively related to one word, the lure word, that is not presented on the list. At test, participants are instructed to recall only the words from the study list, but falsely recall the lure word with high probability (in some cases the lure word is recalled more often than list words). Results of this kind have led to the development of dual-route memory models where the verbatim level information supports accurate recall whereas the gist level information that is activated by the semantic organization of the list supports the intrusion of the lure word (Brainerd *et al.*, 1999; Brainerd *et al.*, 2002). These models were designed to measure the relative contribution of gist and verbatim information in memory but do not provide a computational account for how the gist and verbatim information is encoded in memory.

To explain how the dual-route topic model accounts for the false memory effect, we applied the model to a recall experiment by Robinson and Roediger (1997). In this experiment, each study list contains a number of words that are associatively related to the lure word, which itself is not presented on the study list. The remaining words were random filler words that did not have any obvious associative structure. In the experiment, the number of associatively related words were varied while keeping the total number of study words constant. Figure 15.5(a) shows some example lists that contain 3, 6, and 9 associates of the word ANGER which itself is not present on the list. Figure 15.5(b) shows the observed recall probabilities for the studied items and the lure word as a function of the number of associates on the list. With an increase in the number of associates, the results show an increase in false recall of the lure word

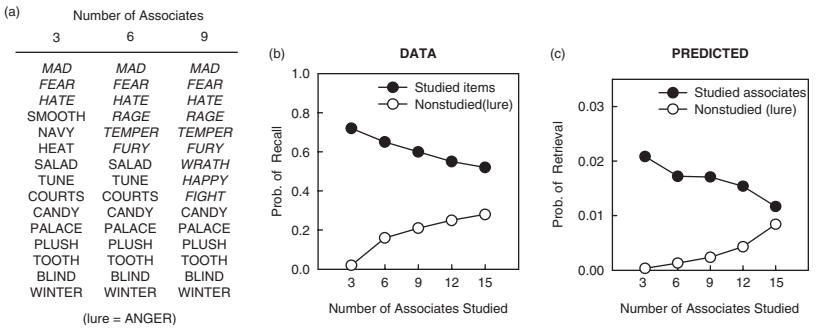

Fig. 15.5. (a) Example study lists varying the number of words associated to the lure ANGER which is not presented on the list. (b) Data from Robinson and Roediger (1997), Experiment 2, showing the observed recall probabilities for studied items and the lure item as a function of the number of associates on the list. (c) Predictions from the dual-route topic model.

and a decrease in veridical recall. We applied the dual-route topic model to this experimental setup and simulated word lists similar to those used by Robinson and Roediger (1997). Figure 15.5(c) shows that model predicts retrieval probabilities that are qualitatively similar to the observed recall probabilities. As the number of associates increases, the model will put increasingly more weight on the topic route, because the topic route can better explain the associative structure when more associates are present. By putting more weight on the topic route, this leads to an increase in generalization beyond the list words, which is associated with an increase in false recall. Similarly, with an increasing weight on the topic route, there is a corresponding decrease in weight for the special-word route. This route is needed to reconstruct the specific words present on a list and as the weight on this route decreases, there is a decrease in veridical recall. Therefore, the model explains these findings in a qualitative fashion by underlying change in the balance between gist and verbatim level information. One advantage of this model over other dual route memory models (e.g., Brainerd *et al.*, 1999; Brainerd *et al.*, 2002) is that the model explains performance at the level of individual words and specifies a representation for gist and verbatim information.

Application to Information Retrieval

The dual-route topic model can be applied to documents to probabilistically decompose words into contextually unique and gist related words. Such as decomposition can be useful for information retrieval because it allows queries to be matched to documents at two levels of generality: specific information captured by the special-word route and content related information captured by the topic model. To illustrate how the model operates on documents, we applied the model with $T = 100$ topics to a set of 1281 abstracts from *Psychological Review*, and separately to a set of 3,104 articles from the *New York Times*. Figure 15.6 shows fragments of two example documents

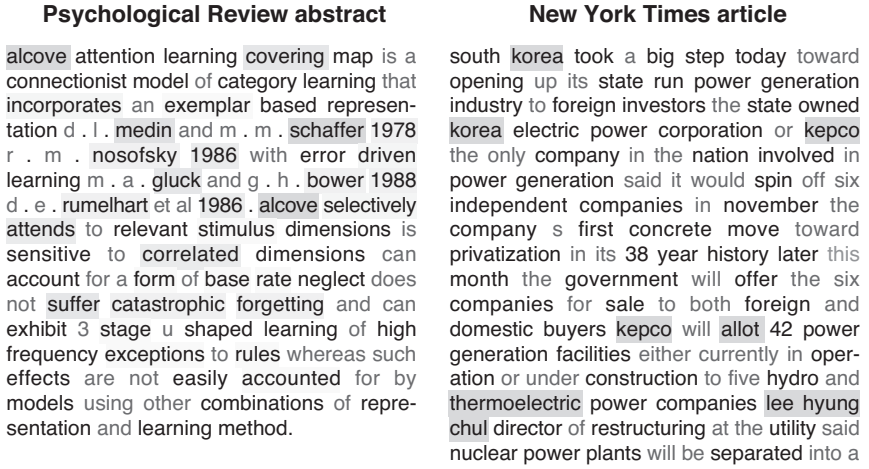

Fig. 15.6. Finding contextually unique words in two example documents. The background shading indicates the probability that a word is assigned to the special-word route.

that were encoded with the dual-route topic model. The background color of words indicates the probability of assigning words to the special words topic – darker colors are associated with higher probability that a word was assigned to the special topic. The words with gray foreground colors were treated as stopwords and were not included in the analysis. The model generally treats contextually unique words as special words. This includes names of people (e.g., NOSOFSKY, SCHAFFER in the psych review abstract) and low frequency words (e.g., THERMOELECTRIC in the New York Times article).

Chemudugunta, Smyth and Steyvers (2007) reported some initial information retrieval results of the dual-route topic model. They applied the model to a several sets of articles from the TREC corpus, which was developed by the information retrieval community to compare and test methods. For each candidate document, they calculated how likely the query q was when 'generated' from the distributions associated with topics and special words. Under the assumption that the query words are generated independently, the query likelihood can be calculated by:

$$P(q \mid d) = \prod_{w \in q} \left[P(x = 0 \mid d) \sum_{t=1}^{T} P(w \mid z = t) P(z = t \mid d) + P(x = 1 \mid d) P'(w \mid d) \right] \tag{5}$$

where the product is over all words that are part of the query. The retrieval performance of the model can be assessed by comparing the query likelihoods to human relevance judgments that are part of the TREC database. Chemudugunta *et al.* (2007) showed that the dual-route topic model significantly outperforms a variety of information retrieval methods such as LSI and LDA which focus on content-based matching and TF-IDF which focuses on keyword matching.

The results of this test indicate that the dual-route topic model does not suffer from the weakness of techniques such as LSI and LDA, which are not able to match specific words in queries and therefore might be prone to over-generalization. Similarly, the model does not suffer from the limitations of the TF-IDF approach in terms of its ability to generalize. The results thus suggest that the best information retrieval results can be obtained by a combination of content-based and keyword-based matching techniques, paralleling contemporary accounts of the structure of human memory.

Discussion

In a rational analysis of cognition, the cognitive system is analyzed in terms of the computational demands that arise from the interaction with our environment (Anderson, 1990; Chater & Oaksford, 1999; Marr, 1982; Oaksford & Chater, 1998). We proposed that both human memory and internet search faces similar computational demands. Both systems attempt to retrieve the most relevant items from a large information repository in response to external cues or queries. This suggests not only that there are many useful analogies between human memory and internet search but also that computational approaches developed in one field potentially lead to novel insights in the other.

For example, we have shown how the PageRank algorithm, developed for the Google search engines to rank webpages, can be useful in understanding human retrieval from semantic memory. We showed how PageRank can be used to measure the prominence of words in a semantic network by analyzing the associative link structure between words. The PageRank measure outperforms other measures for prominence such as word frequency in predicting performance in a simple fluency task. We also showed how research in memory that distinguishes between verbatim and gist information can lead to new computational approaches for encoding and retrieval that are not only useful to explain phenomena such as isolation and false memory effects related to human memory, but can also lead to new information retrieval methods. The central idea in these methods is striking the right balance between content-based (i.e., gist) and form-based (i.e. verbatim) matching approaches when comparing the query to candidate documents.

There are exciting new possibilities for cognitive research in language and memory to influence the design of search engines. If the user formulates a query to a search engine, this query is likely to be influenced by a complex combination of memory and language processes. The user is unlikely to remember all the details of a particular document that needs to be retrieved and therefore cognitive theories of memory organization, encoding, retention and retrieval become relevant. Similarly, the content that is indexed by search engines is often produced by human activity that can be described and explained from a cognitive perspective. While it should not be surprising that there are many cognitive aspects to information retrieval (e.g., Spink & Cole, 2005), often such cognitive aspects are stated quite informally based on intuitive notions of user behavior. For example, in the original paper that motivated the Google search engine, Brin and Page (1998, p. 108) mentioned that the PageRank algorithm was specifically designed as a measure of importance because it

'corresponds well with people's subjective ideas of importance'. Cognitive research can help to formalize and empirically validate intuitive notions of user behavior and the representation and usage of information in memory. Therefore, the connection between cognitive and information retrieval research can work in both directions.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*, 814–823.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*, 396–408.

Anderson, J. R., & Schooler, L. J. (2000). The adaptive nature of memory. In E. Tulving & F. I. M. Craik (Eds.) *Handbook of memory* (pp. 557—570). New York: Oxford University Press.

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General, 128*, 32–55.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Brainerd, C. J., Reyna, V. F., & Mojardin, A. H. (1999). Conjoint recognition. *Psychological Review, 106*, 160–179.

Brainerd, C. J., Wright, R., & Reyna, V. F. (2002). Dual-retrieval processes in free and associative recall. *Journal of Memory and Language, 46*, 120–152.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*, 107–117.

Bruce, D., & Gaines, M. T. (1976). Tests of an organizational hypothesis of isolation effects in free recall. *Journal of Verbal Learning and Verbal Behavior, 15*, 59–72.

Burrell, Q.L. (1980). A simple stochastic model for library loans. *Journal of Documentation, 36*, 115–132.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science, 3*, 57–65.

Chemudugunta, C., Smyth, P., & Steyvers, M. (2007). Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In: *Advances in Neural Information Processing Systems*, 19.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review, 82*, 407–428.

Croft, W. B., & Lafferty, J. (Eds.) (2003). *Language modeling for information retrieval*. Kluwer Academic Publishers.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990) Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41(6)*, 391–407.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17–22.

Dunlosky, J., Hunt, R. R., & Clark, A. (2000). Is perceptual salience needed in explanations of the isolation effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 649–657.

Fabiani, M., & Donchin, E. (1995). Encoding processes and memory organization: A model of the von Restorff effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 224–240.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*, 211–245.

Green, R. T. (1956). Surprise as a factor in the von Restorff effect. *Journal of Experimental Psychology, 52*, 340–344.

Griffiths, T. L., and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science, 101*, 5228–5235.

Griffiths, T. L., Steyvers, M., & Firl, A (in press). Google and the mind: predicting fluency with PageRank. *Psychological Science.*

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic association. *Psychological Review, 114*, 211–244.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 354–384.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. Psychological Science, 17, 767—773.

Hofmann, T. (1999) Probabilistic latent semantic indexing. In *Proc. 22nd Intl. Conf. Res. Dev. Inf. Retriev.* (SIGIR'99) (pp. 50–57). ACM.

Hunt, R. R., & Lamb, C. A. (2001). What causes the isolation effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(6), 1359–1366.

Jenkins, W. O., & Postman, L. (1948). Isolation and spread of effect in serial learning. *American Journal of Psychology, 61*, 214–221.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163–182.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Lezak, M. D. (1995). *Neurological assessment* (3rd ed.). New York: Oxford University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*, 252–271.

McEvoy, C. L., Nelson, D. L., & Komatsu, T. (1999). What's the connection between true and false memories: The different roles of inter-item associations in recall and recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*, 1177–1194.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Nelson, D. L., Dyrdal, G., & Goodmon, L. (2005). What is preexisting strength? predicting free association probabilities, similarity ratings, and cued recall probabilities. *Psychonomic Bulletin & Review, 12*, 711–719.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The university of south Florida word association, rhyme, and word fragment norms. (http://www.usf.edu/FreeAssociation/).

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608–631.

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition.* Oxford: Oxford University Press.

Plaut, D., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56–115.

Ponte, J. M. & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of ACM-SIGIR*, 275–281.

Potter, M. C. (1993). Very short term conceptual memory. *Memory & Cognition, 21*, 156–161.

Robinson, K. J., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science, 8*(3), 231–237.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin and Review, 8*, 385–407.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: MacGraw-Hill.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review, 4*, 145–166.

Seidenberg, S. M., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523–568.

Spink, A., & Cole, C. (Eds.) (2005) *New Directions in Cognitive Information Retrieval*. Springer.

Steyvers, M., Griffiths, T.L. (2006). Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (Eds.), *Latent Semantic Analysis: A Road to Meaning*. Mahwah, NJ: Erlbaum.

Steyvers, M., Griffiths, T.L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences, 10*(7), 327–334.

Steyvers, M., Tenenbaum, J., Wagenmakers, E.J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27*, 453-489.

von Restorff, H. (1933). Uber die Wirkung von Bereichsbildungen im Spurenfeld [the effects of field formation in the trace field], *Psychologische Forschung, 18*, 299–342.