PAI: Automatic Indexing for Extracting Asserted Keywords from a Document

Naohiro Matsumura
PRESTO, JST
The University of Tokyo
Tokyo, 113–8656 Japan
matumura@miv.t.u-tokyo.ac.jp

Yukio Ohsawa
PRESTO, JST
University of Tsukuba
Tokyo, 112–0012 Japan
osawa@gssm.otsuka.tsukuba.ac.jp

Mitsuru Ishizuka
The University of Tokyo
Tokyo, 113–8656 Japan
ishizuka@miv.t.u-tokyo.ac.jp

Introduction

With the increasing number of electronic documents, automatic indexing from a document is an essential approach in information retrieval systems, i.e., search engines. Over the years there have been many suggestions as to what kind of features contribute to an index for the retrieval of documents. For example, the number of occurrences of terms [1] in a document, known as TF (Term Frequency), is considered to be a useful measurement of term significance (Luhn 1957). The number of occurrences of terms over the document collection, known as IDF (Inverse Document Frequency), is also a useful measurement (Spark-Jones 1972). TFIDF, the production of TF and IDF, is used for measuring the discrimination of a document from the remainder of the document collection (Salton & McGill 1983). TF and TFIDF are tend to strongly regard frequent terms as significant. On the other hand, some researches are focused on the lowest-frequent term extraction (Weeber, Vos, & Baayen 2000). Heuristics for the location of terms (e.g., terms in titles and headlines are important) (Baxendale 1958), and for cue terms (e.g., 'final' suggests the start of conclusion) (Edmundson 1969) are also used for detecting the importance of terms.

These stochastic or heuristic measurements are widely used in document retrieval. However, in order to retrieve documents matching users' specific and unique interests, the traditional methods of approach mentioned above are insufficient in that they often disregard the author's specific and original point (Ohsawa, Benson, & Yachida 1999). KeyGraph (Ohsawa, Benson, & Yachida 1999) focuses on extracting keywords representing the asserted main point in a document. The strategy is that the author's main point is based on the fundamental concepts represented by the co-occurrence between frequent terms in a document. We expand the idea of KeyGraph by considering the term activities together with the story of a document.

This paper proposes an automatic indexing method called PAI (Priming Activation Indexing) that extracts keywords representing the author's main point from a document based on the priming effect in cognitive process. The basic idea of PAI is that since an author writes a document emphasizing his/her main point, impressive terms born in the mind of the reader could represent the asserted keywords. PAI employs a spreading activation model without using corpus, thesaurus, syntactic analysis, dependency relations between terms, or any other knowledge except for stop-word list. Experimental evaluations are reported by applying PAI to journal/conference papers.

Priming Effect

Most of cognitive process involving the understanding/interpreting of a document is still little understood. However, the mechanism of memorization in the reader's mind empirically comes out. The human mind can be modeled as a network where concepts are connected to a number of other concepts and the states of concepts are expressed by the activities. If a concept is activated, its adjacent concepts are in turn activated. Thus, activities spread through the network. Many experiments indicate that the speed of associating a concept is in proportion to the level of activity. This kind of phenomenon is known as *priming effect* (Lorch 1982; Balota & Lorch 1986). For example, if 'bread' is activated, 'butter' is named/recognized faster than other unrelated terms.

The priming effect is considered to be closely related to the process of understanding/interpreting a document in the reader's mind. Usually, an author emphasizes his/her main point in the document content, and we go on understanding/interpreting by activating related concepts as we read the content. Here, we define the author's main point as follow.

Definition 1 *Activated terms in the reader's mind represent the author's main point in the document.*

Based on Definition 1, we regard highly activated terms as strongly memorized terms in the reader's mind, and extract them as keywords representing the author's main point.

Spreading of Activation

Spreading Activation Model

The mechanism of human mind, i.e., priming effect at understanding/interpreting a document, has been formalized as *Spreading Activation Model* based on the empirical experiments in cognitive science (Quillian 1968; Collins & Loftus 1975; Anderson 1983). In this model, terms are represented

[1] In this paper, we call a word/phrase as a term.

as nodes, and relations between the terms are represented as associative links between the nodes. In this paper, We call the network as *activation network*.

The activities of nodes propagate along the links to connected nodes. Highly activated nodes are enhanced for further cognitive process. The activity level is determined by the frequency and recentness of activating (Anderson 1995). One of the mathematical formalization of spreading activation model, on which our approach is based, is described as follows (Pirolli, Pitkow, & Rao 1996).

$$\mathbf{A}(t) = \mathbf{C} + ((1-\gamma)\mathbf{I} + \alpha\mathbf{R})\,\mathbf{A}(t-1) \qquad (1)$$

Where, $\mathbf{A}(t)$ is a vector represents the activities of nodes at discrete step $t = 1, 2, \cdots, N$, where $\mathbf{A}(t)_i$ represents the activity of node i at step t. $\mathbf{R}$ is a matrix representing activation network, where $\mathbf{R}_{i,j}\ (i \neq j)$ represents the strength of association between node i and j, and the diagonal elements $\mathbf{R}_{i,j}\ (i = j)$ contains zeros. $\mathbf{C}$ is a vector that represents the activities pumped into the activation network $\mathbf{R}$, where $\mathbf{C}_i$ represents the activities pumped in by node i. $\mathbf{I}$ is an identity matrix. $\gamma\ (0 < \gamma < 1)$ is a parameter for relaxing the node activity, and α is a parameter for determining the amount of activities from a node to its neighbors.

Eq. (1) supposes the situation where the activation network $\mathbf{R}$ is stable regardless of step t. However, in the case of reading a document, it is natural for us to consider that the activation network changes as the story flows because a document has a story through which the author builds his/her arguments. In our view, the flow of activation strongly derived from the story can be a key for understanding the author's specific and original point. The pumped activities $\mathbf{C}$ can be ignored because it is already included in activation network. Accordingly, we transform the spreading activation model in eq. (1) into the following, by replacing $\mathbf{R}$ with $\mathbf{R}(\mathbf{t})$ representing activation network at step t, and setting $\mathbf{C} = 0$.

$$\mathbf{A}(t) = ((1-\gamma)\mathbf{I} + \alpha\mathbf{R}(t))\,\mathbf{A}(t-1) \qquad (2)$$

This translation is an expansion of spreading activation model in eq. (1) for understanding author's main point.

Activation Network

Activation network $\mathbf{R}(t)$ stands for the association between terms in the reader's mind at step t. Here we assume that $\mathbf{R}(t)$ corresponds to the concept of semantically coherent sentences within a document, e.g., sentences in a section/subsection. We call each portion as a *segment*. In reading a document, the author's main point is interpreted by activating $\mathbf{R}(t)$ in turn.

We construct the association between terms in each segment by calculating the co-occurrence of the terms proposed in (Ohsawa, Benson, & Yachida 1999). The algorithm is based on the assumption that associated terms tend to occur within the same sentence. The outline process to a segment is as follows. First, certain terms are extracted as fundamental concepts. Then, the association between the terms are calculated, and links are built between them.

PAI: Priming Activation Indexing

Pre-processing

In advance, three pre-processes are conducted to facilitate and improve the analysis of a document. The most frequent terms, e.g., 'a' and 'it', are considered to be common and meaningless (Luhn 1957). For this reason, we first remove *stop words* used in the SMART system (Salton & McGill 1983). Second, based on the assumption that terms with a common stem usually have similar meanings, various suffixes -ED, -ING, -ION, -IONS are removed to produce the stem word. For example, SHOW, SHOWS, SHOWED, SHOWING are translated into SHOW. In PAI, we employ Porter's suffix stripping algorithm (Porter 1980). Suffix stripping is sometimes an over-simplification since words with the same stem often mean different things in different contexts. However, PAI deals with the problem of understanding the context by spreading the activities along the story of a document. Third, the sequences of terms in a document are recognized as phrases (Cohen 1995).

The Algorithm of PAI

The algorithm of PAI consists of five steps.

Step1) Pre-processing: In preparation, remove stop words, strip suffix, and recognize phrases from a document.

Step2) Segmentation: According to the semantic coherency, a document is segmented into portions $S_t\ (t = 1, 2, \cdots, n)$.

Step3) Activation network: For each segment $S_t\ (t = 1, 2, \cdots, n)$, terms are sorted by their frequencies, and top N%[2] terms are denoted by $K(t)$ as fundamental concepts. The association of terms w_i and w_j is defined as

$$assoc(w_i, w_j) = \sum_{s \in S_t} \min(|w_i|_s, |w_j|_s), \qquad (3)$$

where $|x|_s$ denotes the count of x in sentence s. Pairs of terms in $K(t)$ are sorted by *assoc*, and the pairs above the (*number of terms in* $K(t)$) - 1 th tightest association are linked (Ohsawa, Benson, & Yachida 1999). In addition, we also consider the following factors:

- Priming effect becomes strong in proportion to the strength of association between terms.
- The activation value from w_i is equally divided by the number of links connected to w_i.

For links between w_i and w_j, $\mathbf{R}(t)_{i,j}$ is defined as

$$\mathbf{R}(t)_{ij} = \frac{assoc(w_i, w_j)}{links(w_i)},$$

where $links(w_i)$ denotes the number of links connected to w_i. Other element in $\mathbf{R}(t)$ is defined as 0.

Step4) Spreading activation: From S_1 to S_n, activities are propagated by iterating eq. (2). Primal activity of each term before executing spreading activation is 1. The parameters of γ and α have to be set by trial and error because they depend on the characteristics of documents.

[2] Empirically, we set N as 20.

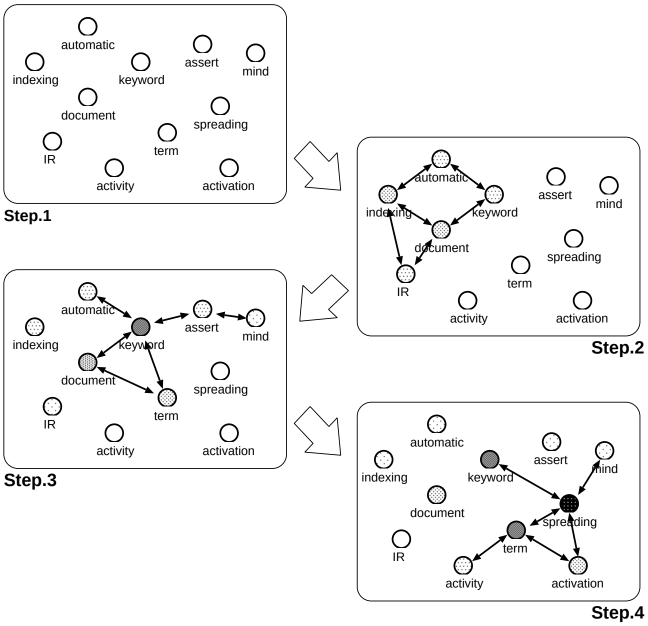

Figure 1: The process of PAI.

Step5) Extract keywords: After spreading activation on all the segments in turn, highly activated terms are considered as the author's main point. However, even if the activity is not so high, a term connecting fundamental concepts is also considered as the author's point (Ohsawa, Benson, & Yachida 1999). As fundamental concepts propagate a large number of activity into neighbors, the activity of a term connecting fundamental concepts can be recognized by focusing on the activity for its frequency of activation. For this reason, we extract both *highly activated terms* and *keenly activated terms* as author's main point.

An Example of PAI

Here we show an example of PAI process. Figure 1 illustrates the transitions of term activities while reading the abstract of this paper. Spreading activation process goes on from Step 1 to Step 4 in turn. The darkness of a node in Figure 1 shows the level of term activity.

Step.1 shows the initial state of the reader's mind. In this state, all terms have equally low activities, e.g., 1. In the first state of reading the abstract, the left-hand terms in Step.2 construct an activation network, and 'automatic', 'indexing', 'keyword', 'document', and 'IR' are activated. On further reading of the abstract, the upper- and right-hand terms in Step.3 reconstruct an activation network, in which the activities of Step.2 come. In the final state, the lower- and right-hand terms in Step.4 reconstruct an activation network and activate the terms as well. The state of Step.4 shows the level of activities of the reader's mind after reading the abstract. From here, we extract highly/keenly activated terms, such as 'spreading', 'activation', 'term', 'activity', 'keyword' etc. as keywords representing the author's main point.

Experimental Evaluations and Discussions

Segments and Parameters

Hereafter, we treat a journal/conference paper as a document. The paper usually consists of several sections/subsections. Each content has semantically coherent context. Therefore, we segment a paper by section/subsection. As for the parameter γ, we assume that the author of a paper does not consider the reader's forgetfulness although the activity of the reader's mind decrease over time (Tanenhaus, Leiman, & Seidenberg 1979). According to the assumption, we set $\gamma = 0$ so as not to decrease term activities during the reading of a document. As for the parameter α, we cannot have any assumption in advance because $\mathbf{R}(t)$ affected by α is derived from various assumptions. In this paper, we determine $\alpha = 1$ by preliminary experiments done before formal experiments.

Case Study

Let us show an output of PAI. The paper (Matsumura, Ohsawa, & Ishizuka 2000) we analyze here describes a new approach of information retrieval for satisfying a user's novel question by combining related documents. The extracted keywords by PAI, TF, TFIDF and KeyGraph are shown in Table 1, and the activation network is shown in Figure 2. The corpus for TFIDF is constructed from 166 papers obtained from Journal of Artificial Intelligence Research [3].

According to the author's comments, the most important terms are 'combination retrieval' and 'document set' ('multiple documents' is also used in the same meaning). It is not a surprise that all methods highly rank 'combination retrieval' (KeyGraph ranks it at 13th) because the term is the most frequent term in the paper. However, 'document set' obtained by PAI cannot be extracted by the other methods. In addition, 'meaning context', 'conditional term', 'abductive inference', 'small number', 'minimal cost', 'past question' are retrieved only by PAI although they also represent the author's main point.

In TFIDF, a term with high DF value is hard to be obtained even if it is significant. For example, TFIDF regards 'abductive inference' as insignificant because it often occurs in the field of Artificial Intelligence. In addition, it is hard to be obtained by TF because the frequency of 'abductive inference' is low.

The advantage of PAI that can extract keywords representing the author's main point regardless of the frequency is derived from the strategy of spreading activation and document segmentation. In the paper, 'abductive inference' is described as extracting 'document set' by 'combination retrieval'. For this reason, the activity of 'abductive inference' becomes high due to the activities of 'document set' and 'combination retrieval'. KeyGraph also makes use of co-occurrence of terms to understand the author's main point, however, the graph is rather perspective than PAI.

Experimental Evaluation

To evaluate the performance of PAI, we compared the keywords obtained by PAI, TF, TFIDF, and KeyGraph. 6 sub-

[3] http://www.cs.washington.edu/research/jair/

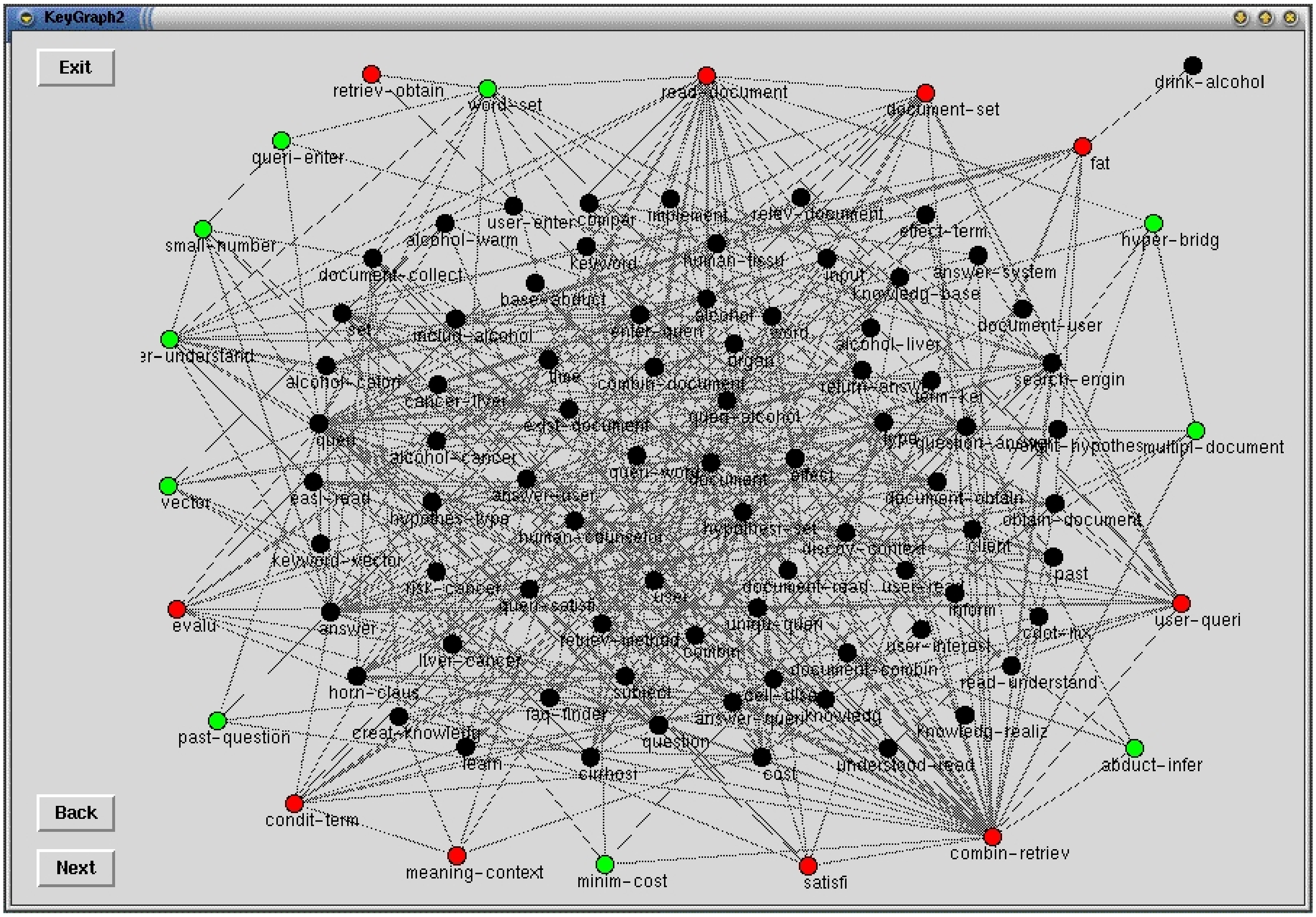

Figure 2: Activation network in a paper (Matsumura, Ohsawa, & Ishizuka 2000). The figure depicts the network in each segment together. The gray nodes denote the keywords extracted by PAI. You can see ‘multi-document’ (right-hand), ‘document-set’ (upper right-hand), ‘combin-retriev’, ‘abduct-infer’, ‘past-question’ (lower right-hand), ‘small-number’ (upper left-hand), ‘meaning-context’, ‘condit-term’ (lower left-hand), ‘minim-cost’ (lower hand).

jects participated in our experiments. From the subjects, we collected 23 journal/conference papers written by each subject. Experiments were conducted as follows: First, from each paper, we extracted 15 keywords by PAI, TF, TFIDF, and KeyGraph individually. Here we regarded the keywords of PAI as top 10 highly activated terms and top 5 keenly activated terms. Then, let each author evaluate each keyword extracted from his own papers to see whether it matches his assertion or not.

Precision (how many of the keywords relevant to the author’s main point are obtained) and *recall* (how many of the retrieved keywords are relevant to the author’s main point) are traditionally used to evaluate information retrieval effectiveness. In our experiment, however, *recall* can not be efficiently computed because the keywords representing the author’s main point cannot be fully extracted even by the author. Instead, we use *mean frequency* of keywords matching author’s main point to evaluate the frequency.

The results of *precision* and *mean frequency* are shown in Table 2. The results show that PAI could extract lower frequency terms more efficiently compared to other keyword extraction methods, despite having almost the same *precision* as TF without corpus. In general, the product of the frequency of terms and the rank order is approximately constant (known as Zipf’s Law (Zipf 1949)). Moreover, infrequent terms are usually insignificant (Luhn 1957). That is, discovering infrequent but significant terms is quite difficult problem. Considering these situations, we can conclude that PAI is a method for extracting infrequent but significant keywords.

Table 2: Experimental results.

	PAI	*TF*	*TFIDF*	*KeyGraph*
precision	0.56	0.55	0.63	0.45
mean frequency	14.3	24.1	19.4	17.9

Table 1: Top 10 keywords obtained by PAI, TF, TFIDF, and KeyGraph.

Ranking	*PAI*†	*PAI*‡	*TF*	*TFIDF*	*KeyGraph*
1	user queri	abduct infer	combin retriev	combin retriev	document
2	read document	small number	document	document	alcohol
3	fat	user understand	user	queri	user
4	satisfi	minim cost	queri	user	query
5	evalu	multipl document	answer	answer	doc
6	retriev obtain	queri enter	knowledge	read document	weights
7	document set	vector	obtain	alcohol	subject
8	meaning context	word set	word	keyword	fat
9	condit term	hyper bridg	read document	question answer	understandable
10	combin retriev	past question	alcohol	answer queri	types

†: highly activated keywords ‡: keenly activated keywords

Conclusion

Because an author writes a document emphasizing his/her specific and original point, impressive terms born in the mind of the reader could represent the author's main point. Based on this assumption, we proposed PAI which realizes priming effect in the reader's mind for keyword extraction. Experimental evaluation shows that PAI can extract keywords representing the author's main point regardless of the frequency.

Chance discovery is defined as the awareness on and the explanation of the significance of a chance, especially if the chance is rare and its significance is unnoticed (Ohsawa 2002). From this point of view, PAI can be a tool for supporting chance discovery because understanding asserted keywords leads us aware of the significance of the document.

References

Anderson, J. 1983. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior* 22:261–295.

Anderson, J. 1995. *Cognitive psychology and its implications*. Freeman, 4 edition.

Balota, D., and Lorch, R. 1986. Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, Cognition* 12:336–345.

Baxendale, P. 1958. Man made index for technical literature - an experiment. *IBM Journal of Research and Development* 2(4):254–361.

Cohen, J. 1995. Highlights: Language- and domain-independent automatic indexing terms for abstracting. *Journal of American Society for Information Science* 46:162–174.

Collins, A., and Loftus, E. 1975. A spreading-activation theory of semantic processing. *Psychological Review* 82:407–428.

Edmundson, H. 1969. New methods in automatic abstracting. *Journal of ACM* 16(2):264–285.

Lorch, R. 1982. Priming and searching processes in semantic memory: A test of three models of spreading activation. *Journal of Verbal Learning and Verbal Behavior* 21:468–492.

Luhn, H. 1957. A statistical approach to the mechanized encoding and searching of literary information. *IBM Journal of Research and Development* 1(4):309–317.

Matsumura, N.; Ohsawa, Y.; and Ishizuka, M. 2000. Combination retrieval for creating knowledge from sparse document collection. In *Proceeding of Discovery Science*, 320–324.

Ohsawa, Y.; Benson, N. E.; and Yachida, M. 1999. Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. 12–18.

Ohsawa, Y. 2002. Chance discoveries for making decisions in complex real world. 20(2).

Pirolli, P.; Pitkow, J.; and Rao, R. 1996. Silk from a sow's ear: Extracting usable structures from the web. In *Proceeding of CHI*, 118–125.

Porter, M. 1980. An algorithm for suffix stripping. *Automated Library and Informations Systems* 14(3):130–137.

Quillian, M. 1968. *Semantic Memory, Semantic Information Processing*. MIT Press.

Salton, G., and McGill, M. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Spark-Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(5):111–121.

Tanenhaus, M.; Leiman, J.; and Seidenberg, M. 1979. Evidence for multiple stages in the processing of ambiguous words in syntactiv contexts. *Journal of Verbal Learning and Verbal Behavior* 18:427–440.

Weeber, M.; Vos, R.; and Baayen, R. 2000. Extracting the lowest-frequency words: Pitfalls and possibilities. *Computational Linguistics* 26(3):301–317.

Zipf, G. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.