From Search Engines to Question-Answering Systems: The Problems of World Knowledge, Relevance, Deduction, and Precisiation

Lotfi A. Zadeh*

Summary. Existing search engines, with Google at the top, have many truly remarkable capabilities. Furthermore, constant progress is being made in improving their performance. But what is not widely recognized is that there is a basic capability which existing search engines do not have: deduction capability – the capability to synthesize an answer to a query by drawing on bodies of information which reside in various parts of the knowledge base. By definition, a question-answering system, or a Q/A system for short, is a system which has deduction capability. Can a search engine be upgraded to a question-answering system through the use of existing tools – tools which are based on bivalent logic and probability theory? A view which is articulated in the following is that the answer is: no.

The first obstacle is world knowledge – the knowledge which humans acquire through experience, communication, and education. Simple examples are: "Icy roads are slippery," "Princeton usually means Princeton University," "Paris is the capital of France," and "There are no honest politicians." World knowledge plays a central role in search, assessment of relevance and deduction. The problem with world knowledge is that it is, for the most part, perception-based. Perceptions – and especially perceptions of probabilities – are intrinsically imprecise, reflecting the fact that human sensory organs, and ultimately the brain, have a bounded ability to resolve detail and store information. Imprecision of perceptions stands in the way of using conventional techniques – techniques which are based on bivalent logic and probability theory – to deal with perception-based information. A further complication is that much of world knowledge is negative knowledge in the sense that it relates to what is impossible and/or nonexistent. For example, "A person cannot have two fathers," and "Netherlands has no mountains."

The second obstacle centers on the concept of relevance. There is an extensive literature on relevance, and every search engine deals with relevance in its own way, some at a high level of sophistication. But what is quite obvious is that the problem of assessment of relevance is quite complex and far from solution.

There are two kinds of relevance (a) question relevance, and (b) topic relevance. Both are matters of degree. For example, on a very basic level, if the question is q: "Number of cars in California?" and the available information is p: "Population of California is 37,000,000," then what is the degree of relevance of p to q? Another example: To what degree is a paper entitled "A New Approach to Natural Language Understanding" of relevance to the topic of machine translation.

Basically, there are two ways of approaching assessment of relevance (a) semantic, and (b) statistical. To illustrate, in the number of cars example, relevance of p to q is a matter of semantics and world knowledge. In existing search engines, relevance is largely a matter of statistics, involving counts of links and words, with little if any consideration of semantics. Assessment of semantic relevance presents difficult problems whose solutions lie beyond the reach of bivalent logic and probability theory. What should be noted is that assessment of topic relevance is more amendable to the use of statistical techniques, which explains why existing search engines are much better at assessment of topic relevance than question relevance.

The third obstacle is deduction from perception-based information. As a basic example, assume that the question is q: What is the average height of Swedes?, and the available information is p: Most adult Swedes are tall. Another example is: Usually Robert returns from work at about 6 p.m. What is the probability that Robert is at home at 6:15 p.m.? Neither bivalent logic nor probability theory provide effective tools for dealing with problems of this type. The difficulty is centered on deduction from premises which are both uncertain and imprecise.

Underlying the problems of world knowledge, relevance, and deduction is a very basic problem – the problem of natural language understanding. Much of world knowledge and web knowledge is expressed in a natural language. A natural language is basically a system for describing perceptions. Since perceptions are intrinsically imprecise, so are natural languages.

A prerequisite to mechanization of question-answering is mechanization of natural language understanding, and a prerequisite to mechanization of natural language understanding is precisiation of meaning of concepts and proposition drawn from a natural language. To deal effectively with world knowledge, relevance, deduction, and precisiation, new tools are needed. The principal new tools are: precisiated natural language (PNL); protoform theory (PFT), and the generalized theory of uncertainty (GTU). These tools are drawn from fuzzy logic – a logic in which everything is, or is allowed to be, a matter of degree.

The centerpiece of the new tools is the concept of a generalized constraint. The importance of the concept of a generalized constraint derives from the fact that in PNL and GTU it serves as a basis for generalizing the universally accepted view that information is statistical in nature. More specifically, the point of departure in PNL and GTU is the fundamental premise that, in general, information is representable as a system of generalized constraints,

with statistical information constituting a special case. This, much more general, view of information is needed to deal effectively with world knowledge, relevance, deduction, precisiation, and related problems.

In summary, the principal objectives of this paper are (a) to make a case for the view that a quantum jump in search engine IQ cannot be achieved through the use of methods based on bivalent logic and probability theory; and (b) to introduce and outline a collection of nonstandard concepts, ideas, and tools which are needed to achieve a quantum jump in search engine IQ.