Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 79 (2013) 4 – 17

Procedia
Social and Behavioral Sciences

9th Conference on Applications of Social Network Analysis (ASNA)

Semantic Network Analysis as a Method for Visual Text Analytics

Philipp Drieger[a,*]

[a]*noumentalia.de - digital arts, science & technology, Grünauer Str. 82, 86633 Neuburg a. d. Donau, Germany*

Abstract

This paper proposes an approach on a method for visual text analytics to support knowledge building, analytical reasoning and explorative analysis. For this purpose we use semantic network models that are automatically retrieved from unstructured text data using a parametric *k*-next-neighborhood model. Semantic networks are analyzed with methods of network analysis to gain quantitative and qualitative insights. Quantitative metrics can support the qualitative analysis and exploration of semantic structures. We discuss theoretical presuppositions regarding the text modeling with semantic networks to provide a basis for subsequent semantic network analysis. By presenting a systematic overview of basic network elements and their qualitative meaning for semantic network analysis, we describe exploration strategies that can support analysts to make sense of a given network. As a proof of concept, we illustrate the proposed method by an exemplary analysis of a wikipedia article using a visual text analytics system that leverages semantic network visualization for exploration and analysis.

Keywords: semantic network analysis; network visualization; visual text analytics; method for text analysis

1. Introduction

The concept of semantic networks has a long history (Quillian, 1968) and opened up a basis for knowledge modeling and representation (Helbig, 2006) by providing an adaptable formal framework for scientific developments and applications. Semantic networks allow to model semantic relationships (Sowa, 1991) that are represented in a graph with labeled nodes and edges. Graph theory (Diestel, 2005) and network analysis (Brandes, 2005; Newman, 2010) allow to process and analyze networks formally. With a focus on structural properties, a rich spectrum of methods and metrics evolved with social network analysis (Wassermann & Faust, 1994) and network sciences (Newman et al., 2006) to analyze and characterize network structures. Additionally, graph drawing (Di Battista, 1999) offers advantages for the visualization of networks and for visual exploration and analysis (von Landesberger et al., 2010). Visual analytic approaches combine automated data analysis with interactive visualizations (Keim et al., 2008b) to support analytic processes (Keim et al., 2008a) which involve analytical reasoning and knowledge building. In the light of visual analytics, semantic networks can automatically be retrieved from unstructured text data to be used as a medium for visual text analytics. Automated relation extraction from text

[*] http://www.noumentalia.de/
E-mail address: philipp.drieger@noumentalia.de

doi:10.1016/j.sbspro.2013.05.053

(Batagelj et al., 2002; Diesner & Carley, 2004) involves information retrieval and text mining techniques (Feldman & Sanger, 2007; Risch et al., 2008; Berry & Kogan, 2010). The seamless interaction of an analyst with the retrieved network requires an interactive system that helps to interpret semantics as a ”system of signs” (Loebner, 2002).

Our proposed methodological approach for *semantic network analysis* builds upon a visual text analytics system (Drieger, 2012) that allows for a user-centered visual exploration and analysis of complex semantic networks. By folding linear text into a network of interconnected words, this system provides an alternative view on text and enables the analyst to explore the context of semantic structures. Compared to the sole linear reading of a text, the system helps to gain analytical insights from the given semantic network. The interpretation of semantic structures is supported by the interplay of quantitative and qualitative analysis of the network structure. Motivated by the goal to support an analyst in general text analytic tasks, we present basic elements of a method for visual text analytics that adapts semantic networks.

2. Related work

Related work can be found in different areas that are relevant to our approach. We draw a focus on text mining, natural language processing, knowledge representation, network analysis and visual analytics. *Text mining* techniques are used for automated information retrieval from textual data sources (Feldman & Sanger, 2007; Berry & Kogan, 2010) according to a given task and text model. Large-scale document analysis mostly builds on word frequency based vectorization and dimensionality reduction for representation (Risch et al., 2008), e.g. self-organizing maps (Kohonen et al., 2000), techniques based on multidimensional scaling, principal component analysis (Jolliffe, 2002) or latent semantic indexing (Deerwester et al., 1990). Information retrieval from natural language texts can be improved and refined by using *natural language processing* (NLP) techniques (Manning et al., 2008). NLP allows for better tokenization, e.g. by using stemming, thesauri or shallow parsing to provide preciser models of natural language text that can be used for relation extraction (e. g. (Brandes & Corman, 2002)). We partially build on text mining techniques (Feldman & Sanger, 2007) but as we follow a structural approach first (see section 3), we try to use a minimum of NLP. Thus, we concentrate on a text model that builds on windowing techniques using a *k*-next-neighborhood model for relation extraction (see section 4). In future works this model can be refined with NLP or machine learning techniques on demand. To further improve relation extraction and model building, encyclopedic knowledge representations like WordNet (Miller, 1995) or SALSA (Burchardt et al., 2006) may be adopted for concept mining.

Knowledge representation techniques are also related to our approach which is based on generalized semantic networks (Sowa, 1991). In contrast to models that make use of formalized schemes for knowledge representation (Helbig, 2006), we propose to model a semantic network in close connection to the empirical text data. As we do not rely on ontologies, our model provides direct access to the text source and is not restricted to classifications that depend on specific ontologies. Thus, we cannot provide logical reasoning like SNePS (Shapiro, 1999) or classifications according to a given ontology. We rather rely on lightweight annotation techniques (Drieger, 2012) to enhance a given text model with meta data which is expressed in natural language or involves tagging like folksonomies (Mika, 2005) in the sense of ”dynamic ontologies” (Sowa, 2006). Frame semantics (Fillmore, 1982; Busse, 2012) or approaches on cognitive modeling for declarative knowledge representation (Anderson, 1980) may also be related to our method, but we first draw a focus on network exploration and *network analysis* referring to (Diestel, 2005; Brandes, 2005; Newman et al., 2006). By including methods of social network analysis (Wassermann & Faust, 1994; INSNA, 2012), we can build on a rich framework to analyze semantic network structures to refine and adopt them for semantic analysis which is mainly focussed on qualitative aspects.

Using a *visual text analytics* system that allows to apply network analysis to automatically retrieved semantic networks and provides interactive visualization, our approach is closely related to visual text analytics. Visual text analytic systems (e.g. SPIRE (Wise et al., 1995)) rely on text visualization to encode different analytic properties (Risch et al., 2008). Systems like DocuBurst (Collins et al., 2009) or TextArc (Paley, 2002) help to visually summarize a given text. Other visualization types focus on pattern recognition (Wattenberg, 2002) or feature extraction. Contextual informations are mostly represented in tree structures (Lee et al., 2006; Wattenberg & Viégas, 2008). Phrase Nets (van Ham et al., 2009) allows to build a more complex semantic network with user defined word relations to map unstructured text. We deem the visual analytics approach (Keim et al., 2008a) appropriate for the visual exploration and analysis of complex graphs (von Landesberger et al., 2010) to ease the recognition of semantic relations and structurally relevant patterns by leveraging interactive visualization for exploration.

3. Theoretical presuppositions

Finding an appropriate model for text can be a demanding task in text mining and information retrieval. According to general model theory (Stachowiak, 1973), models are *representations of objects* that capture *selected properties* and also depend on *pragmatic goals* of the model's creator. Models can be seen as intentional constructions of an observer (Von Foerster, 1984) who relies on certain construction rules to perceive some object. Thus, creating a model for text greatly depends on the pragmatic aspects and its purposes according to the user, especially in the context of semantics (Loebner, 2002).

Depending on these assumptions we consider a minimal model to approach semantics in unstructured text. Referring to de Saussure (de Saussure, 2011) and Eco (Eco, 1986), we take up the general attempts of structuralism and semiotics to understand text as a structure that can be formalized as a system of signs: ”A sign constitutes a sign only as a part of a system – only insofar as it is related to, and different from, the other signs of the language.” (Loebner, 2002) By modeling text as a system of signs which is represented by differently related words, we obtain a formal structure of word differences and relations which corresponds to a graph $G = (V, E)$ with $E \subseteq [V]^2$ according to graph theory (Diestel, 2005). V is a set of nodes that represents different words as unique keys and E is a set of edges that stand for relations between the different words to capture a given system of signs in a text. Thus, G is a formal model of a generalized semantic network (Sowa, 1991) that consists of different words being related to each other. Word relations are represented as undirected edges and are not specified by a formal ontology (Helbig, 2006). This approach of modeling text as an interconnected system of signs is closely related to the concept of a ”rhizome” (Deleuze & Guattari, 1987) which can be considered as a structural metaphor for natural language as every word can potentially be connected to any other word due to some meaning. Referring to the topological structure of a rhizome, Eco (Eco, 1986) postulated that ”Model Q”, proposed by Quillian (Quillian, 1968), would be an appropriate descriptive model for the semantics of natural language which is rather characterized as a ”maze of language” than some well-ordered system. If language is a possible representation of knowledge (Helbig, 2006), text artifacts can provide an empirical basis of communication. Considering Serres' theory of communication which also recurs to the formal structure of a network (Serres, 1968) we can use networks to model communications that are expressed in text artifacts.

On the basis of these theories, we use the proposed type of a semantic network to model text as a network of connected words. Textual linearity is folded into a non-linear network of signs that can be arranged spatially by using network layout algorithms. We deem this kind of a semantic network to be a possible model of an empirical text source as it encodes semantics in word differences and their relations by a given textual structure. As one word can refer to more than one meaning, different word senses are symbolized in a *semantic field* which is represented by node adjacencies in our model. As a set of words encodes a semantic field that mostly denotes some complex meaning, this model comprises the meanings with respect to their interconnected semantic properties. Using this approach, we follow the paradigm to use natural language as a given structure of knowledge representation that can be analyzed according to a chosen model. One goal is to provide a basic method to explore text sources for knowledge building and to analyze them with the help of *semantic network analysis* (see section 5) to support human interpretation. Thus, our approach may contribute to the spectrum of methods in visual text analytics and applied network analysis that draw a focus on semantics.

4. Automated network retrieval

According to our theoretical presuppositions we model unstructured text with a generalized semantic network model (Sowa, 1991) by using relation extraction techniques similar to (Danowski, 1993; Batagelj et al., 2002; Diesner & Carley, 2004). As we focus on relational in-depth analysis we construct a semantic model from word collocations (Loebner, 2002). For this purpose we tokenize different words in sentences and apply a stoplist to filter semantically irrelevant words (Feldman & Sanger, 2007). Word relations are built on a k-next-neighborhood model with user-adjustable k, so that every word is connected with its k predecessors and k successors in a sentence. The retrieved network yields an undirected, weighted graph consisting of nodes which represent words, and edges that correspond to the extracted word relations. Nodes are attributed with text origins to have a reference to the underlying source text. Weights are calculated by applying quantitative measurements (e.g. centrality measures) to affect the layout of the network. The layout is computed by using an adjustable spring-embedding algorithm (Di Battista, 1999) to obtain a spatial representation that is used for exploration and visual analysis. Similar to other dimensionality reduction techniques, we encode semantic similarity with spatial proximity. For our initial studies we choose this basic model of relation extraction which can further be refined with more advanced NLP techniques like full speech parsing for domain-specific analysis on demand.

A major problem of analyzing unstructured text data concerns the ambiguity of natural language (e.g. synonymy and polysemy) and semantic complexity (Loebner, 2002). Another problem for automated data analysis is the lack of explicit knowledge of contextual and situational information (Helbig, 2006). Moreover, we cannot assume that a given text source is grammatically correct or uses precise syntax and terminology. Often text sources are heterogeneous and include slang, abbreviations, metaphors or intentional semantics that cannot be resolved without contextual information which mostly depends on commonsense knowledge about the extensional world. Therefore, applying strict ontologies to generally classify arbitrary text sources remains a challenge. Although natural language processing is highly elaborated, we try to use as few NLP as possible in the first step to concentrate on a structural approach with a simple text model. Our approach is further motivated by the pragmatic goal to support an analyst with a methodological framework and a tool to explore and analyze an arbitrary text source by using a general model for representing text in an interactive system which allows for intuitive exploration and analysis of textual data.

Table 1. Semantic network analysis and exploration for qualitative and quantitative interpretation of semantic network elements

Element	Quantity	Quality
Network structure	Statistics like degree distribution	Characteristic topological properties of the network.
Edges	Weight	Collocation, semantic relation, meaning.
Paths	Length	Set of connected semantic relations.
Nodes	Degree ($d(v)$)	Complexity of a semantic concept on word basis. Position in a semantic field.
Hubs	Centrality measures, filtering ($d(v) \geq n$)	*Global* importance of a node relative to the network. Node position in a *local* context, e.g. connection between complex semantic concepts or topics.
Subgraphs	(see network structure above)	Complex semantic context encoded in interconnected semantic fields.
Clusters	Clustering coefficient; filtering ($d(v) \leq n$)	Strongly connected components encoding specific semantic topics or complex concepts.

5. Semantic Network Analysis

Given the model of a semantic network and an interactive system for visual analysis and exploration, we propose methodological elements for semantic network analysis. In our terms *semantic network analysis* describes the process of analyzing the structure of a semantic network (see figure 2) that is retrieved according to the given text model. Therefore, we suggest to distinguish semantic network analysis in automated network analysis that yields quantitative measurements and human-centered analysis of semantics that help to reveal qualitative aspects of a semantic network. *Automated network analysis* uses algorithms to calculate metrics and statistics of a given

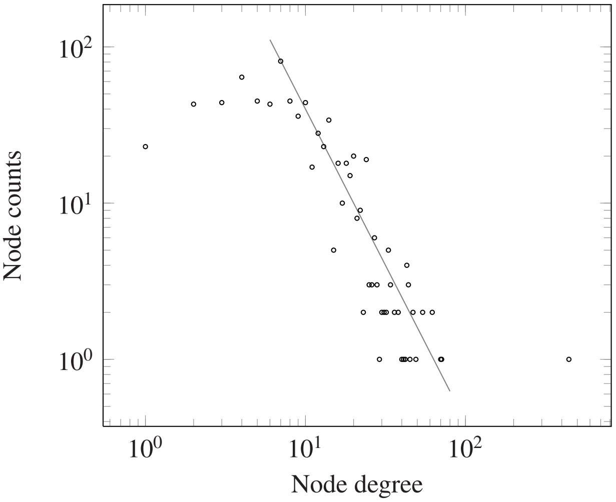

Fig. 1. Diagram of the degree distribution for the exemplary semantic network N_1 with $|V| = 747, |E| = 4370$. This network was automatically retrieved from the wikipedia article about ”knowledge” (Wikipedia, 2012a) as described in section 4. The control line has slope $\lambda = 2.00$ indicating a power law distribution $d(v) \sim k^{-\lambda}$ according to (Barabasi & Albert, 1999). The global hub ”knowledge” can be identified as an outlier with the highest node degree.

network (Wassermann & Faust, 1994; Brandes, 2005; Newman et al., 2006; Newman, 2010), thus yielding *quantitative* measurements that characterize the structure of a network and can also support the process of qualitative semantic analysis. *Semantic analysis* is mainly based on the process of human interpretation and understanding of semantic structures for the purpose of exploration or analytical reasoning. This process is supported by domain-specific and commonsense knowledge, mainly the knowledge about the extensional world (Helbig, 2006) to discover *qualitative* aspects from a given semantic network. We draw this distinction because neither an analyst is capable of fast automated processing of large amounts of data nor is a computer able to explore and analyze semantics by truly understanding extensional meaning. Thus, both types of analysis are to complement each other for *semantic network analysis* as stated in this work. Upon this distinction, table 1 provides a systematic overview of quantitative and qualitative aspects of basic elements for semantic network analysis, referring to graph theory (Diestel, 2005), methods of network analysis (Brandes, 2005), social network analysis (Wassermann & Faust, 1994) and network science (Newman et al., 2006). In case of special analytic tasks, more advanced methods, metrics and techniques can be added on demand. The following subsections discuss basic elements in semantic network analysis. A *network* structure consists of *nodes* and *edges*. Nodes have a position in the network and can be characterized as *global or local hubs*. *Subgraphs* are parts of a network structure, e.g. the neighborhood of a node. *Clusters* are subgraphs that contain strongly connected nodes. On the basis of these elements we first discuss their use for quantitative and qualitative semantic network analysis. As a proof of concept we demonstrate the application of the proposed method in an exemplary text analysis of a wikipedia article (see section 6).

5.1. Network structure

The overall structure of a given network can be examined quantitatively by using statistical methods. Figure 1 shows the node distribution for our example network which follows a power law (Barabasi & Albert, 1999). The power law distribution helps to characterize the topology of the network and provides hints for the analysis of nodes and clusters. As nodes encode words, this distribution is correlated to Zipf’s law (Zipf, 1949) as shown in (Masucci & Rodgers, 2006). This topological property can also be confirmed by studies about scale free conceptual networks retrieved from language (Motter et al., 2002) and the small-world property of human language (i Cancho & Solé, 2001). Knowing about these properties, the analyst can easily determine important nodes or explore local clustering. Measurements like the diameter or density help to characterize a network structure in comparison with other networks. Many metrics can be aggregated and used for network layout or more advanced

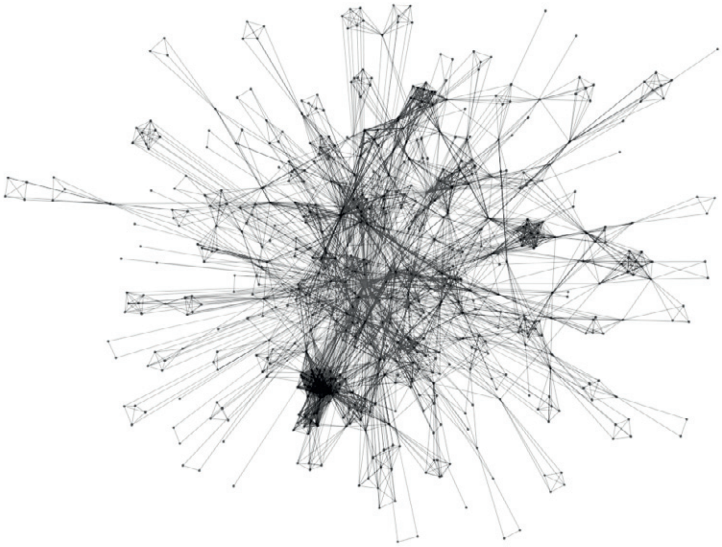

Fig. 2. Overview of the semantic network retrieved from (Wikipedia, 2012a). Local hubs (see figure 3) and clusters (see section 5.5) are grouped around the central hub ”knowledge”. The topology can be characterized as a scale-free network (see figure 1) with small-world properties (see section 5.1).

measurements and statistics.

Respecting the density, we can estimate if a network is sparse or dense. Qualitatively, sparse networks can indicate a lack of global semantic coherency on word basis. Instead, dense networks that show a high density may indicate higher amounts of actualized word relations and appear to be more coherent. The amount of nodes can indicate the qualitative variety of present words in semantic fields and thus a higher variety in language expression. The qualitative exploration of a larger semantic network structure can help to indicate salient substructures. Hubs and clusters can be recognized and contextualized in the whole structure. By examining the overall structure of a network we obtain a global overview. Figure 2 shows our example network which clearly shows highly connected nodes and clusters.

5.2. *Edges*

Edges represent relations between two nodes and may be weighted according to statistical quantities of adjacent nodes like their centrality measures. In our model relations are undirected edges and correspond to word collocations according to the chosen model of automated network retrieval. Qualitatively, edges indicate that two words are related to each other by word collocation. As we are not relying on an ontology that could be applied to the given network model, the semantic specification has to be explored by the analyst. The given model supports the analyst to find the relevant relations. As an edge is defined by two nodes, the analysis of a semantic relation mostly involves the comparison of two nodes respecting their adjacencies as shown in figure 6. Edges may also be interpreted in more complex subgraphs (see section 5.6) or paths.

5.3. *Paths*

A path describes a set of connected edges that connect two nodes with each other. Every path has a length which serves as a quantitative characterization. Paths can be used to determine possible (shortest) connections between two nodes which help to find out how nodes are connected to each other. If a path exists between two nodes we can state for a semantic network that there is either a direct or an indirect semantic relation on word basis. According to the k-next neighborhood model direct semantic relations correspond to sentences and indirect

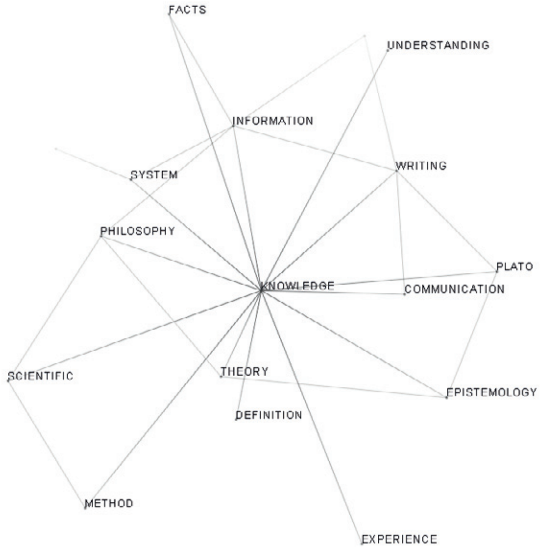

Fig. 3. Semantic topics represented by highly connected nodes after subgraph construction with degree threshold $d(v) \geq 18$. The retrieved nodes are local hubs in the overall structure (see figure 2) and provide access to semantic concepts that are further encoded in local clusters (see figure 5).

relations represent second order semantic relations due to multiple sentences that have words in common. Paths can be analyzed qualitatively to reveal different semantic relations that connect words in the semantic network. If there is no existing path between two words the analyst could postulate that there are missing semantic relations. Paths may also be useful to describe analytical points of interest in a concrete analysis (see section 6.2).

5.4. Nodes

Nodes can quantitatively be characterized by degree which indicates the amount of adjacent nodes and measures a node's connectedness. As nodes represent words, the amount of adjacent nodes also provides qualitative insights in the diversity of a word's usage referring to the position in the underlying semantic network model. Words with a high degree imply a highly differentiated relatedness and typically indicate local hubs in a semantic network (see section 5.5). Given a point of interest, a node can be used as a starting point to explore its adjacent nodes to examine the given semantic field of word collocations (see figure 6) to approach meanings. Given a set of nodes which are connected according to their adjacencies, more complex semantic contexts can be obtained for qualitative analysis. To explore the semantic contexts at a given node, the adjacencies of its adjacent nodes are expanded and displayed as a subgraph (see figure 7). Following a hypothesis, the analyst can build a subgraph at some point of interest and investigate the connections found there. He may also attach another word that is not in the set to see how it is connected to the given set. Using this kind of subgraph strategies, nodes can be explored and compared according to the given network structure.

5.5. Hubs

Hubs represent important nodes in a network and often correspond to highly connected nodes. Beside the node degree, other centrality measures (e.g. betweenness or eigenvalue centrality) can be employed to quantitatively characterize the importance of a node in a network (Wassermann & Faust, 1994; Brandes, 2005). Qualitatively, *global hubs* indicate central nodes in a semantic network that have multiple connections to other nodes with low degree, thus indicating frequently referred core issues. *Local hubs* represent abstract concepts that are frequently referenced from different contexts. By excluding stopwords from the network, local hubs can reveal connecting

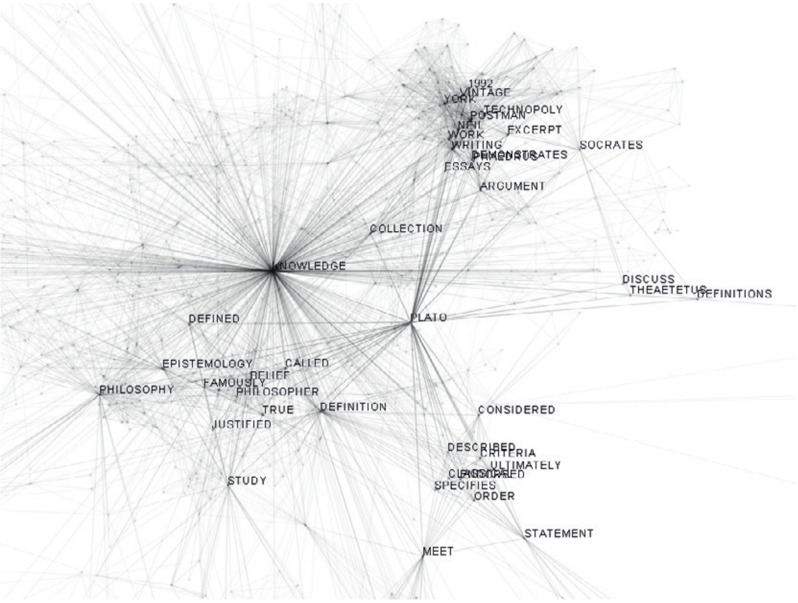

Fig. 4. Selection of a node at some point of interest (”plato”) to examine the position and connectedness in the network. ”Plato” can be identified as a local hub that is connected to at least four clusters that encode specific semantic fields. Compared to figure 5, ”plato” has a higher influence in the network than ”kant”.

topics in a semantic network and can be obtained by degree filtering using a threshold t to obtain a subgraph G’ = (V’, E’) where V' satisfies $\{v \in V | d(v) \geq t\}$. The retrieved subgraph G’ contains highly connected nodes which provide an abstract summary of the text source. Local hubs may also act as a broker, connecting between more complex semantic concepts that include local word clustering (see figure 5). Figure 3 shows semantically high related topics in our example which clearly reveals an abstract summary of the article referring to the most important words.

5.6. Subgraphs

Subgraphs describe parts of a given network according to a chosen subset of nodes and edges. A subgraph can easily be constructed at some point of interest by expanding the adjacent nodes of a selected node. This strategy can also be applied to a set of selected nodes that are to be analyzed. Of course, other subgraphs can be constructed arbitrarily or by node or edge filtering (e.g. degree thresholds or n-cores). Qualitatively, subgraphs represent semantic contexts as nodes are arranged according to their local connections. Figure 7 illustrates semantic contexts at two selected nodes according to their adjacencies. The analysis of semantic contexts provides information how nodes are locally embedded in the whole network and help the analyst to find related information.

5.7. Clusters

Clusters are subgraphs consisting of strongly connected components which can be measured with a node-based clustering coefficient (Watts & Strogatz, 1998). Alternatively, filtering nodes with lower degree $d(v) \leq n$ can reveal clusters as cliques. Qualitatively, clusters represent groups of strongly connected words and thus may indicate complex semantic concepts that are described by words which aren’t likely to be connected to other semantic concepts. Thus, clusters help to indicate concepts that represent special topics. As local hubs are mostly located between clusters to connect them as brokers, clusters and hubs complement each other in the interpretation. With this, we can analyze how topics are connected by local hubs. Referring to (i Cancho et al., 2004; Ravasz & Barabasi, 2003), language networks show local clustering and a recursive hierarchical organization which correspond to characteristic topological properties as stated in section 5.1 and can be helpful for explorative analysis.

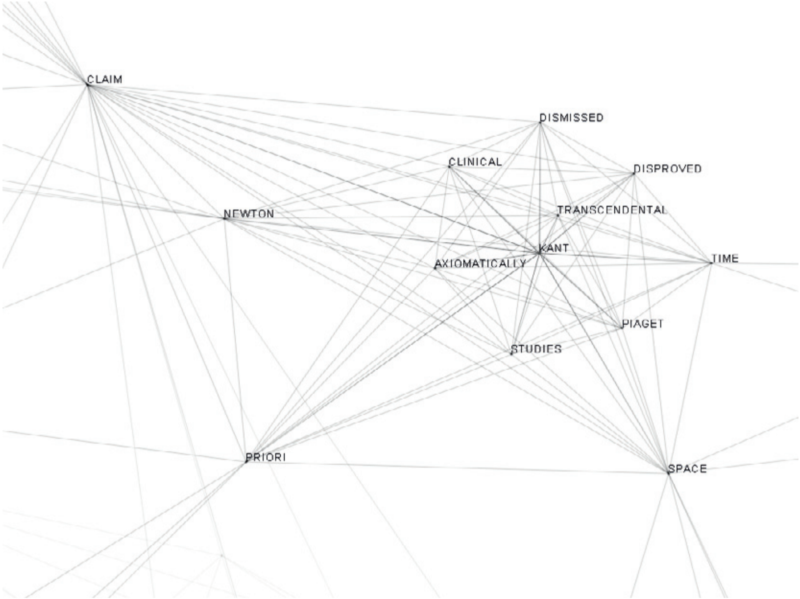

Fig. 5. Local clustering at the node ”kant” with connecting hubs in the periphery. Clusters represent local semantic fields that can encode special topics, as its nodes have no significant connections to other nodes.

6. Exemplary application

According to the given basic elements of semantic network analysis, we can deduct elements of a method for the exploration and analysis of semantic networks. The analysis of a network structure provides *topological insights*. This overview may be used as a starting point for *detailed exploration and navigation* at some point of interest. If the graph is too large, filtered graphs can be used for overview. With a focus on nodes, they are characterized as global or local hubs, connecting brokers or isolates according to their position in the network structure. Adjacent edges can be explored to *investigate semantic fields and contexts* in subgraphs that show the neighborhood at a node. The complementary analysis of clusters and local hubs reveals different *semantic topics and issues* and show how they are connected with each other. Nodes are contextually related in the light of their adjacencies or connecting paths. With this, parts of a network can be *compared* on the basis of nodes and edges according to their position in the network structure.

The following sections illustrate the analysis of an exemplary semantic network according to the elements of semantic network analysis given in section 5. With respect to the interplay of quantitative and qualitative aspects we analyze the wikipedia article about ”knowledge” (Wikipedia, 2012a) by applying the proposed methodological elements from section 5. We utilize our text visualization system (Drieger, 2012) to verify our findings in the visualization.

6.1. Overview

First, we start with figure 2 to examine the *overall structure* of the retrieved network with applied layout. We can clearly identify a dominant *hub* in the center of the structure and significant *local clustering*, suggesting a scale-free topology. The *degree distribution* confirms the *global hub* as an outlier and shows a power law distribution (see figure 1) which is typical for the suggested topology. Additionally, *local clusters* are connected to *local hubs* again. Qualitatively, we can state that the text is strongly focussed on a core issue which is represented by the central node which acts as a *global hub*. This core issue is distinctively surrounded by different *local hubs* that are connected to *local clusters* of descriptions using specific terminologies due to different key words. Those *local hubs* may be used as starting points to explore different associated semantic fields and contexts in *subgraphs*.

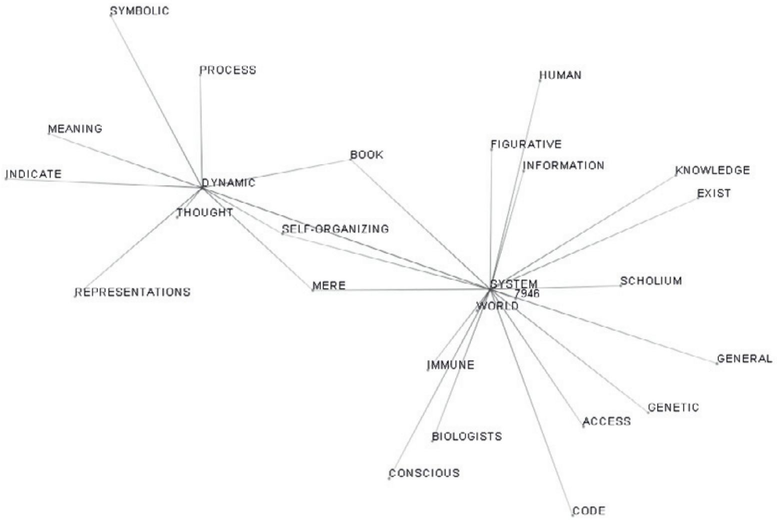

Fig. 6. Relation between the two nodes ”system” and ”dynamic”. Adjacent nodes (1-neighborhood) reveal common nodes ”mere”, ”book” and ”self-organizing” and show the semantic fields around ”system” and ”dynamic”.

6.2. *Topic exploration with local hubs*

In a second step, we try to get an overview about the smaller *local hubs*. Referring to figure 3, we clearly see *local hubs* being connected to the central *hub* node which is identified as the core issue of the article ”knowledge”. The connected *hubs* reveal core topics which are connected to the main issue. Starting from the central *node* we first reach a circular *path* P_1 that connects ”theory”, ”epistemology”, ”plato”, ”writing”, ”information” and ”philosophy” with ”knowledge”, right next to the other directly connected nodes ”definition”, ”experience”, ”communication”, ”facts”, ”system”, ”scientific”, ”method” and ”understanding”, which are partially connected to P_1, too. Although the *path* P_1 doesn’t reveal concrete semantic concepts yet, we get an abstract impression about frequently related topics. For example, we can state at this point that knowledge is related with theories, one of them is identified as epistemology. If we know the meaning of ”epistemology”, we can successfully close this semantic triangle of ”knowledge”, ”theory” and ”epistemology” by formulating the statement that epistemology is a theory about knowledge. If we don’t know anything about these relationships, the underlying source text yields the sentence S_0: ”In philosophy, the study of knowledge is called epistemology, and the philosopher Plato famously defined knowledge as ’justified true belief’” (Wikipedia, 2012a) if we review the first occurrence of ”epistemology” in the source text. The drill-down to the source text also reveals the connection to ”plato” and ”philosophy”. To explore the remaining connections along the *edges* of the given abstract *path* P_1, we can simply retrieve the most relevant core concepts and their relationships to describe ”knowledge”. This is due to the fact of the *network topology* that many other words are recursively connected to those *nodes* in P_1. With respect to quantitative properties of the *network topology*, we can gain insights that are also helpful for qualitative analysis and exploration.

6.3. *Role of a node in the semantic network*

Now we might be interested in the special *node* ”plato” and its role as a *local hub* in the network. As shown in figure 4, ”plato” is connected with at least four *local clusters* that encode different semantic fields. One of them can be matched with the source text given in S_0 above. The others could be examined more closely to investigate the concrete issues. More interestingly, in the light of some common background knowledge, we could raise the question if ”plato” has a greater authority about ”knowledge” in the given *network* compared to ”kant” who was

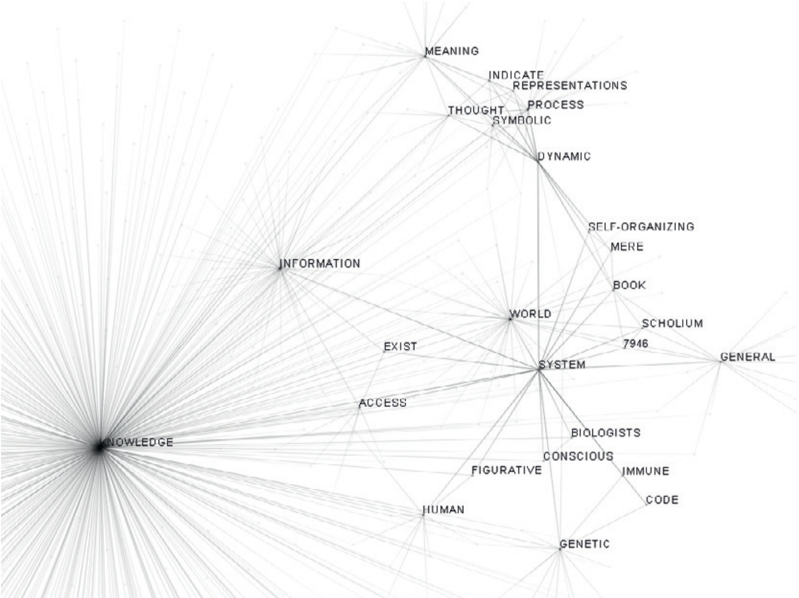

Fig. 7. Subgraph with extended neighborhood (2-neighborhood) at two given nodes of interest (”dynamic”, ”system”) to further explore the semantic contexts in the neighborhood of the subgraph in figure 6.

a very important modern philosopher. If we directly select the *node* ”kant” in the network, we can find it located in a *local cluster* (see figure 5) that turns out to be a rather special topic that is connected to the main topic at greater distance. Thus, we could state the qualitative hypothesis that ”kant” is semantically less important than ”plato” in the context of the given text source. This hypothesis is solely deducted from the analysis of the given *network structure*. We can easily confirm this hypothesis with the given passages in the source text and state that the philosophy of Kant is not well-presented in the given article, e.g. in contrast to the german article about knowledge (Wikipedia, 2012b).

6.4. *Analysis of a more complex question*

In the next step, we assume that we want to analyze a more complex question within the given *network*: ”How is knowledge related to the concept of a dynamic system?” To resolve this question, we first try to find out how ”dynamic” is related to ”system” to see how the compound concept ”dynamic system” is configured in the *network*. Figure 6 shows both nodes with their 1-neighborhood expanded. Obviously they have few *nodes* in common: ”mere”, ”book” and ”self-organizing”. A first source drill-down leads to the sentence S_1: ”The system should apparently be dynamic and self-organizing (unlike a mere book on its own)” (Wikipedia, 2012a) which turns out to be a statement in some context, but no core concept. Nevertheless, we can assume that ”self-organizing” is semantically related to a ”dynamic system”. According to our question, we secondly examine the relation between knowledge and dynamic system by exploring the associated semantic contexts using a *subgraph*. By expanding all *nodes* that are given in figure 6, we obtain the *subgraph* in figure 7 which is more complex due to adjacent interconnections. We can easily analyze that ”dynamic” is situated in less semantic fields than ”system”. Consequently, ”system” appears to be used semantically more differentiated than ”dynamic” which is closely connected to the upper dense *cluster* (see figure 7) covering ”meaning”, ”thought”, ”symbolic”, ”process”, ”representations” and ”indicate”. Instead, ”system” is closely connected to a loose *cluster* covering ”biologists”, ”immune”, ”code”, ”genetic” and ”conscious” which reveals the sentence S_2: ”For biologists, knowledge must be usefully available to the system, though that system need not be conscious.” (Wikipedia, 2012a) If we further examine the text source at both *clusters*, we find that they are not directly related, although they are closely connected to the relation between ”dynamic” and ”system”. Instead, we find that S_1 is denoted as a criteria of S_2 in the source text. According to our question, we can conclude that the most significant relation between

”knowledge” and ”dynamic system” can be found in the context of biological systems. To resolve our question we can formulate the answer from our analyzed facts: ”Biological systems depend on useful knowledge if these are constituted as dynamic, self-organizing systems – whether they are conscious or not”.

7. Conclusions

We discussed a general method for semantic network analysis that covers quantitative and qualitative aspects that arise from the analysis and interpretation of network structures. With regard to the stated theories and the pragmatic aspect of the proposed network model we gave an example for a text analysis along the methodological elements being applied on a wikipedia article. We showed how semantic network analysis can help an analyst to visually explore unknown text sources, develop hypothesis and examine them in detail using analytical reasoning. Neither the proposed method nor the visual text analytics system claim to be a replacement for reading or textual understanding. As a given text source is transformed into a semantic network model, we face an abstraction that provides an alternative perspective on a text source and supports text analysis. Due to the abstraction and the chosen model, we can explore and analyze a text quickly without reading it word by word and see at a glance how words are interconnected and contextually situated in a network structure. By using a visual text analytics system, this can be achieved more efficiently in comparison to a sole reading of the text. Operating on a formalized structure, we can rely on the methodological framework of network analysis and leverage interactive visualization for visual text analytics. This intuitive access provides the advantage to analyze a text apart from its concrete meanings which also determine our interpretation in the reading process. The proposed method can be extended with more advanced methods of network analysis.

Of course, we are facing restrictions that arise from language ambiguity and model complexity. As words may denote many meanings, we face similar difficulties like natural language processing to ensure a proper connection between a given word and its denoted semantic concept. Thus, we cannot exactly map semantic concepts with the chosen model as it remains an approximation due to the abstraction of word collocations. If text sources grow larger, the model complexity increases accordingly. Thus, our approach is not yet applicable to larger text collections. Spurred by these restrictions, we try to further improve our visual text analytics system in future. To obtain clearer and preciser networks we consider to enhance network retrieval with more advanced text mining techniques with natural language processing. Larger document collections can be preprocessed and stored in a graph database that can be used to query subgraphs on demand. The task of preselecting relevant documents can be eased with document classification techniques that make use of the retrieved graph structures. Finally, we are researching on more advanced interactive visualization techniques that allow for a better visual analysis of the retrieved network structures.

Acknowledgments

We would like to thank Andreas Harrer and Michael Weissinger for all discussions and their helpful comments.

References

Anderson, J. R. (1980). *Cognitive psychology and its implications*. A series of books in psychology. San Francisco: Freeman.

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, (pp. 509–512).

Batagelj, V., Mrvary, A., & Zaveršnik, M. (2002). Network analysis of texts. In *T. Erjavec, J. Gros (Eds.), Proc. of the 5th International Multi-Conference Information Society - Language Technologies* (pp. 143–148).

Berry, M. W., & Kogan, J. (2010). *Text Mining: Applications and Theory*. Wiley.

Brandes, U. (2005). *Network analysis: methodological foundations*. Lecture notes in computer science; 3418. Springer, Berlin.

Brandes, U., & Corman, S. (2002). Visual unrolling of network evolution and the analysis of dynamic discourse. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on* (pp. 145–151). doi:10.1109/INFVIS.2002.1173160.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., & Pinkal, M. (2006). The salsa corpus: a german corpus resource for lexical semantics. *Proceedings of LREC, Genoa, Italy*, .

Busse, D. (2012). *Frame-Semantik : ein Kompendium*. Berlin: de Gruyter.

i Cancho, R. F., & Solé, R. V. (2001). The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, *268*, 2261–2266.

i Cancho, R. F., Solé, R. V., & Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical review.E, Statistical, nonlinear, and soft matter physics*, *69*, 051915.

Collins, C., Carpendale, S., & Penn, G. (2009). Docuburst: Visualizing document content using language structure. *Eurographics/ IEEE-VGTC Symposium on Visualization*, .

Danowski, J. A. (1993). Network analysis of message content. In B. G. A. Richards William D. (Ed.), *Progress in communication sciences* (pp. 197–221). Norwood NJ: Ablex volume 12.

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

Deleuze, G., & Guattari, F. (1987). *A Thousand Plateaus: Capitalism and Schizophrenia*. University of Minnesota Press.

Di Battista, G. (1999). *Graph drawing : algorithms for the visualization of graphs*. Prentice Hall, NJ.

Diesner, J., & Carley, K. M. (2004). *AutoMap1.2 - Extract, analyze, represent, and compare mental models from texts*. Technical Report Carnegie Mellon University School of Computer Science ISRI - CASOS.

Diestel, R. (2005). *Graph theory*. Berlin: Springer.

Drieger, P. (2012). Visual text analytics using semantic networks and interactive 3d visualization. In K.Matkovic, & G.Santucci (Eds.), *EuroVA 2012: International Workshop on Visual Analytics* (pp. 43–47). URL: `http://diglib.eg.org/EG/DL/PE/EuroVAST/EuroVA12/043-047.pdf`.

Eco, U. (1986). *Semiotics and the Philosophy of Language*. Indiana University Press.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fillmore, C. J. (1982). Frame semantics. In T. L. S. of Korea (Ed.), *Linguistics in the Morning Calm* (pp. 55–82). Seoul: Hanshin Publishing Co.

van Ham, F., Wattenberg, M., & Viegas, F. B. (2009). Mapping text with phrase nets. *IEEE Transactions on Visualization and Computer Graphics*, *15*, 1169–1176.

Helbig, H. (2006). *Knowledge representation and the semantics of natural language : with 23 tables*. Cognitive technologies. Berlin ; Heidelberg ; New York: Springer.

INSNA (2012). International network for social network analysis. URL: `http://www.insna.org/`.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008a). Visual analytics: Definition, process, and challenges. *Information Visualization: Human-Centered Issues and Perspectives*, (pp. pp. 154–175).

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., & Ziegler, H. (2008b). Visual analytics: Scope and challenges. In S. J. Simoff, M. H. Böhlen, & A. Mazeika (Eds.), *Visual Data Mining* (pp. 76–90). Berlin: Springer.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Paatero, V., & Saarela, A. (2000). Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, *11*, pp. 574–585.

von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., van Wijk, J., Fekete, J.-D., & Fellner, D. (2010). Visual analysis of large graphs. In *Proceedings of EuroGraphics: State of the Art Report*.

Lee, B., Parr, C., Plaisant, C., Bederson, B., Veksler, V., Gray, W., & Kotfila, C. (2006). Treeplus: Interactive exploration of networks with enhanced tree layouts. tvcg.

Loebner, S. (2002). *Understanding semantics*. Understanding language series. London: Arnold.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Masucci, A., & Rodgers, G. (2006). Network properties of written human language. *Physical review.E, Statistical, nonlinear, and soft matter physics*, *74*, 026102.

Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. In *International Semantic Web Conference* (pp. 522–536).

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, *38*, 39–41.

Motter, A. E., de Moura, A. P. S., Lai, Y.-C., & Dasgupta, P. (2002). Topology of the conceptual network of language. *Phys. Rev. E.*, *65*.

Newman, M. (2010). *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc.

Newman, M., Barabási, A.-L., & Watts, D. J. (2006). *The structure and dynamics of networks*. Princeton Univ. Press.

Paley, W. B. (2002). Textarc: Showing word frequency and distribution in text. In *Proc. of the IEEE Symp. on Information Visualization*.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing*. Cambridge.

Ravasz, E., & Barabasi, A. L. (2003). Hierarchical organization in complex networks. *Phys. Rev. E.*, *67(2)*.

Risch, J., Kao, A., Poteet, S. R., & Wu, Y. J. (2008). Text visualization for visual text analytics. In S. J. Simoff, M. H. Böhlen, & A. Mazeika (Eds.), *Visual Data Mining* (pp. 154–171). Berlin: Springer.

de Saussure, F. (2011). *Course in general linguistics*. Columbia University Press.

Serres, M. (1968). *Hermès I. La communication.*. Minuit, Paris.

Shapiro, S. C. (1999). Sneps: A logic for natural language understanding and commonsense reasoning. In L. M. Iwanska, & S. C. Shapiro (Eds.), *Natural Language Processing and Knowledge Representation* (pp. 175–195). MIT Press.

Sowa, J. F. (1991). *Principles of Semantic Networks*. Morgan Kaufmann.

Sowa, J. F. (2006). A dynamic theory of ontology. In *Proceedings of the conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference* (pp. 204–213). Amsterdam: IOS Press.

Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Springer.

Von Foerster, H. (1984). *Observing systems*. The systems inquiry series (2nd ed.). Salinas, Calif.: Intersystems Publ.

Wassermann, S., & Faust, K. (1994). *Social Network Analysis: Methods and Application*. Cambridge, University Press.

Wattenberg, M. (2002). Arc diagrams: Visualizing structure in strings. In *IEEE Symposium on Information Visualization*.

Wattenberg, M., & Viégas, F. B. (2008). The word tree and interactive visual concordance. In *Proc. of the IEEE Conf. on Information Visualization* (pp. 1221–1229). volume 14, 6.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature, 393*, .

Wikipedia (2012a). Knowledge (last visit: April 16, 2012). http://en.wikipedia.org/wiki/Knowledge.

Wikipedia (2012b). Wissen (last visit: April 16, 2012). http://de.wikipedia.org/wiki/Wissen.

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995). Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of the IEEE Information Visualization Symposium InfoViz*.

Zipf, G. (1949). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley Press.