

CONTENT COVERAGE OF ANIMAL BEHAVIOR DATA

Sidharth Thakur^a, Katy Börner^b, Ketan Mane^b, Emilia Martins^c & Terry Ord^c

^aComputer Science Department, Indiana University, Bloomington, IN 47405,

^bSchool of Library and Information Science, Indiana University, Bloomington, IN 47405

^cDepartment of Biology, Indiana University, Bloomington, IN 47405

Abstract

This paper describes the results of analysis and visualization of Animal Behavior research papers published in major journals. Analysis is carried out to obtain associations among the citation records for the domain. The primary goal is to observe the content coverage and the prominent research areas and describe the research activities taking place over the past decade. The methodologies used to study the domain provide a potential platform that can be extended to cover the entire publication history of the domain. The fact that the field of Animal Behavior has remained uncharted to a great extent provides immense motivation for the study.

Keywords: information visualization, latent semantic analysis, knowledge domain visualization, animal behavior

1. INTRODUCTION

The area of Animal Behavior studies is one of the many domains that have witnessed an explosion of research and development activities in the recent years. The sheer volume of research to be reviewed complicates the task of tracing the history of the Animal Behavior domain. This explains why only two studies have attempted such a feat, and both have tried to circumvent this problem by relying on representative measures (i.e., a review of key textbooks [12] or a subset of published research [6]).

Knowledge domain visualizations (KDV) are a special kind of *Information Visualization* [8] that exploit powerful human vision [9, 10] and spatial cognition to help humans mentally organize and electronically access and manage large, complex information spaces [11]. KDV use sophisticated data analysis and visualization techniques to objectively identify major research areas, experts, publications, etc. and their interrelations in a domain of interest. They can be used to gain an overview of a knowledge domain; its homogeneity, import-export factors, and relative speed; to track the emergence and evolution of topics; or to help identify the most productive and innovative research areas. Benefits of KDV include reducing visual search time, revealing hidden relations, displaying data sets from several perspectives simultaneously, facilitating hypothesis formulation, and serving as effective means of communication. In recent years, the study of knowledge domains has equipped researchers, grants committees and experts with a new tool to analyze their domains.

This paper reports a first attempt to study the growth of the Animal Behavior domain and to find the prominent 'hot topics' over the years through the changes in content coverage in the domain. Visualizations are used to provide a more explicit and intuitive representation of the domain to uncover the trend-related information. To support rendering of such plots, graphical visualization software, Pajek [1] was used.

A broad perspective of the growth of the domain can be observed from the patterns of publications among the prominent or 'core' set of journals see Figure 1. This graph shows the increased activities in the area in the last few decades.

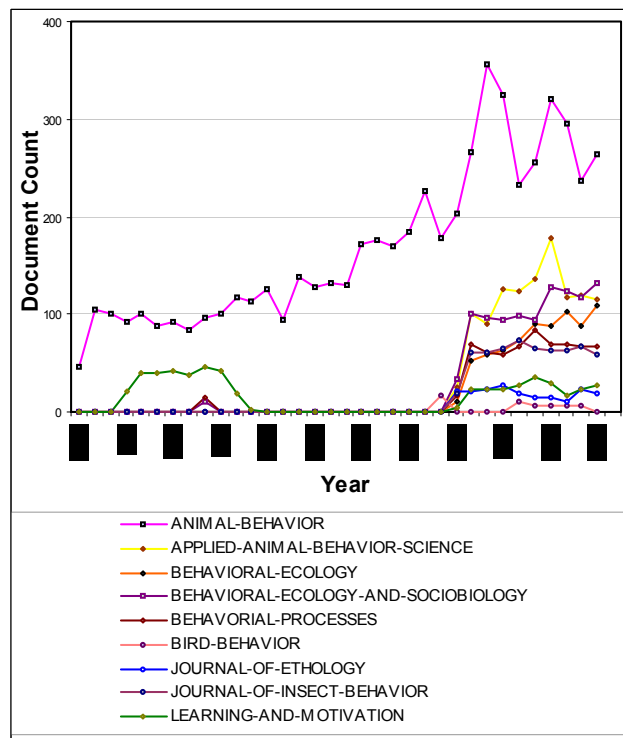


Figure 1: Count of publications in the core journals in the Animal Behavior research domain

The section 2 describes the data set pertaining to the Animal Behavior domain that was used in the study. It describes the various factors that were taken into account to obtain a reliable and workable data set. Section 3 presents the technique and the various methodologies central to the data analysis. The visualizations obtained are discussed in section 4, which are used to describe the domain and discuss the results in section 5. The paper concludes with a discussion on the future work in section 6.

2. DATA SET

The data set used for the study comprises the Animal Behavior citation data obtained from the Biological Abstracts database (published by Biological Abstracts, Inc). The records were obtained for approximately half a million articles appearing in over 200 journals that publish animal behavior research dating back to 1968.

For this study, we focused on the documents belonging to the ‘Core’ set of journals comprising *Animal Behavior*, *Applied Animal Behavior Science*, *Behavioral Ecology*, *Behavioral Ecology and Sociology*, *Behavioral Process*, *Bird Behavior*, *Journal of Ethology*, *Journal of Insect Behavior*, *Learning and Motivation*.

Furthermore, we selected journal articles published in the years 1994, 1997, and 2000 to study the trends over the past decade on a manageable data set. For each document in this dataset, author given keywords used in the journal citations were retrieved. Several documents were found to have zero or just one keyword. Those documents were excluded from the co-word analysis. The year-wise count of these documents is as follows: 1994 (122), 1997 (5) and 2000 (10).

| | 1994 | 1997 | 2000 |
|-----------------------------------|------|------|------|
| Number of documents | 648 | 778 | 740 |
| Number of unique keywords | 1244 | 2324 | 2269 |
| Average number of unique keywords | 1.92 | 2.99 | 3.06 |

From the data in Table 1, it is observed that the average number of keywords increases over time while the number of paper and the number of unique keywords roughly stay the same.

Table 1: Simple statistics for the three data sets

3. JOURNAL COVERAGE

The first part of the analysis tried to identify the major keywords for each core journal. The top ten keyword for each journal in the data sets under consideration are given in Figure 2.

| |
|---|
| <i>JOURNAL OF ANIMAL BEHAVIOR</i> : Behavior, Aggression, Animal-Behavior, Evolution , Male, Female, Sexual Selection , Body-size, Foraging , Reproductive-success, Mate-choice |
| <i>JOURNAL OF BIRD BEHAVIOR</i> : Aggression , Adult, Feeding, Male, Reproductive success , Species interaction, Vocalization, Brood parasitism, Brooding, Brood mates |
| <i>JOURNAL OF BEHAVIORAL ECOLOGY</i> : Behavior, Sexual Selection , Body size, Predation risk, Reproductive success, Mate-choice, Fitness, Parental care, Male, Reproduction |
| <i>JOURNAL OF LEARNING AND MOTIVATION</i> : Behavior, Learning , Neural coordination, Conditioning, Rat, Conditioned stimulus, Motivation , Pavlovian-Conditioning |
| <i>JOURNAL OF APPLIED ANIMAL BEHAVIOR SCIENCE</i> : Behavior, Animal Welfare , Animal-Behavior, Animal Husbandry, Stress, Aggression, Housing, Meeting-Abstract, Social-Behavior, Grazing, Feeding |
| <i>JOURNAL OF INSECT BEHAVIOR</i> : Oviposition, Female, Foraging-Behavior , Body-Size, Sexual Selection , Foraging , Reproduction, Male, Parasitoid |
| <i>JOURNAL OF BEHAVIORAL PROCESS</i> : Behavior, Learning , Abstract, Animal-Behavior, Nervous System, Reinforcement, Social-Behavior, Memory, Female, Foraging-Behavior, Aggression |
| <i>JOURNAL OF BEHAVIORAL ECOLOGY AND SOCIOBIOLOGY</i> : Sexual Selection , Reproductive-Success, Body-Size, Evolution, Female, Aggression, Male, Competition, Reproduction, Sperm Competition |
| <i>JOURNAL OF ETHOLOGY</i> : Female, Aggression , Body-Size, Male, Dominance , Copulation, Spawning, Foraging, Mating-Behavior , Social Behavior |

Figure 2: Top ten keywords per journal

Given in figure 3 are the top ten keywords for each of the three years

In Figure 2, we bolded those keywords, which occur in the list of top ten keywords per year, see Figure 3. Bolded in Figure 3 are those keywords that refer to dominant areas in the three years. These keywords correspond to those identified in the co-Keyword space. Figures 8-10 explain and provide details on those areas.

1994: Behavior, Animal-Behavior, **Animal-Communication**, **Evolution**, Mathematical-Model, Aggression, **Foraging**, Predation, Seasonality, Learning

1997: Behavior, Male, Female, **Reproduction**, Ecology, Adult, Sexual Selection, Animal Husbandry, **Evolution**

2000: **Sexual Selection**, Body-Size, Aggression, Reproductive-Success, Animal-Welfare, Predation-Risk, Territoriality, Mating-System, Mate-Choice, Competition

Figure 3: Top ten keywords per year

4. SEMANTIC DOCUMENT SPACE

Next, we determine the semantic similarity among all documents in each of the 3 years (1994, 1997, 2000). In particular, we are interested in any topic changes in this time span.

Latent semantic analysis (LSA) [5], also called latent semantic indexing, was applied to determine semantic similarity among the documents based on their keywords. LSA extends the vector space model by modeling term-document relationships using a reduced approximation for the column and row space computed by the singular value decomposition of the term by document matrix. The strength of LSA lies in resolving the fundamental issues concerning the conventional lexical matching schemes namely, synonymy (similar meaning words) and polysemy (words with multiple meaning) [3].

Data parsing, generation of unique terms and term vs. document frequency matrices, and similarity matrix computations were carried out using code available in the Information Visualization Repository¹ at Indiana University. The LSA SVDPACKC provided by M. Berry [2] was applied to determine the most important latent dimensions. The most significant dimensions obtained for the three years are: 1994 (114 dimensions), 1997 (112 dimensions) and 2000 (114 dimensions).

Visualizations of the semantic relationships among similar documents were generated using the *Pajek* graph visualization software [1]. The Kamada Kawai algorithm [4] implemented in Pajek was used to layout the documents in a 2-dimensional space.

The results is an overview of the semantic similarity between the various documents in the three data sets and are depicted in Figure 5–7. Each document is represented by a dot. The color of each dot depends on the journal in which papers were published, see Figure 4. Interestingly, documents published in one and the same journal don't cluster. This may indicate that all core journals cover similar topic areas.

Papers with high semantic similarity are placed closer in space and those with low similarity are further apart. Links between document are visible if the similarity value is at least 0.7. The distance between documents does not reflect the exact semantic similarity values of these documents as the layout algorithm aims to map a multi-dimensional similarity space into a two-dimensional layout but also tried to minimize document and link overlap. Transparent circular areas indicate document clusters of interests.

In an attempt to describe the resulting document and keyword spaces, we consulted domain experts to identify major clusters and interpret the most frequently occurring keywords for those clusters. In the 1994 data set three clusters were identified. The first cluster (blue background) deals with documents dealing with parental behavior. The second cluster in red covers animal behavior and learning research. The gray cluster contains documents on feeding behavior.

The 1997 data set contains three main clusters containing documents on aggressive social behavior, sexual selection and mating behavior and sexual behavior.

In 2000 only two clusters were identified: mating and foraging behavior as well as nesting behavior.

5. CO-WORD OCCURRENCE KEYWORD SPACE

Finally, we are interested in the changes of topics covered in the years under consideration. From the previously extracted list of unique keywords, we selected those keywords that occurred at least two times. Subsequently, we determined the similarity among the keywords based on their co-occurrence in the document sets. For example, if BEHAVIOR and MALE are used as keywords in one document then their frequency value is increased by one and hence their similarity increases. The values in the resulting frequency matrices for each year were divided by the

¹ Information Visualization repository: <http://iv.slis.indiana.edu/>

highest value in this matrix. The resulting keyword spaces have been visualized using *Pajek*, using the Fruchterman-Reingold 2D-algorithm [7].

The visualizations are shown in Figure 8-10. Unsurprisingly, BEHAVIOR is the central, highly interconnected node, bridging different characteristics of animals like Aggression, Mating, Welfare in all the three time slices. Phase shift in the research area over the years is denoted by highlighted keywords.

The Figure 8 describes the keyword relations among the most frequently occurring keywords for the year 1994. The most dominant areas of study are identified by: ANIMAL COMMUNICATION, EVOLUTION AND FORAGING. The color-coded regions are used to demarcate a cluster of keywords that represents a common concept. High frequency keywords in these clusters are highlighted.

Similarly, for the year 1997, the demarcated regions identify the prominent research areas namely, REPRODUCTION and EVOLUTION. In 2000, the study of SEXUAL SELECTION dominated the research and the keywords pertaining to this area denoted in green. Some of the co-occurring keywords also describe the study of NATURAL SELECTION, MATING SUCCESS, COURTSHIP, etc. that clearly indicate the relevance to the prime topic of research.

6. DISCUSSION AND FUTURE WORK

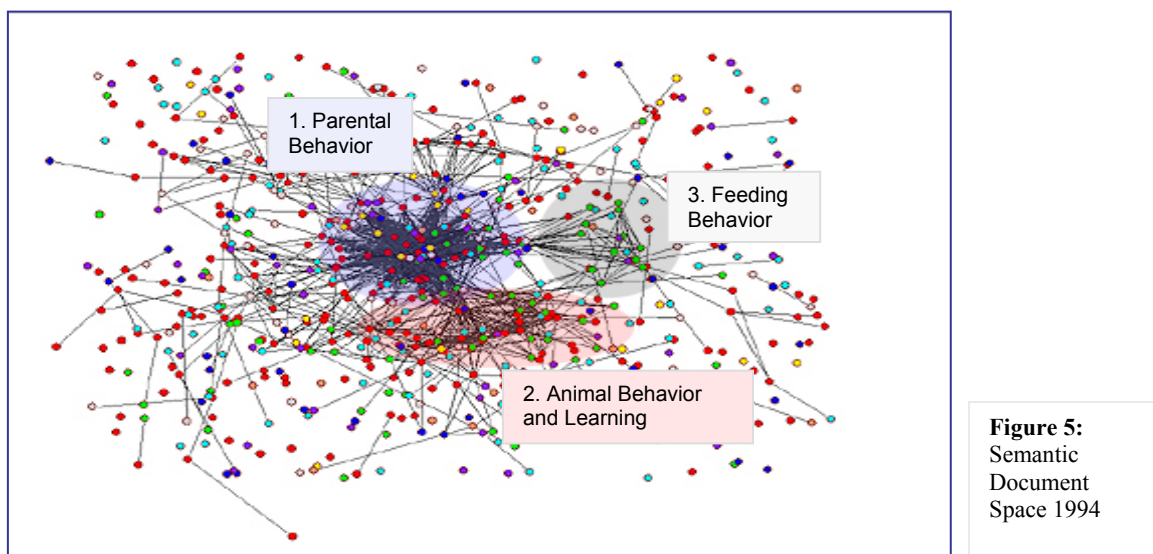
This paper discussed first attempts to map the domain of Animal Behavior research. We showed the coverage of core journals in this area, analyses and visualized the semantic space of documents in three different time slices, and plotted the space of co-occurring keywords for those three years.

The present work focused on the keywords extracted from the corpus of the documents. Using title words would strengthen the document and co-keyword analyses.

Our ultimate goal is to provide a comprehensive overview of the field by considering all material relevant to animal behaviour research. These analyses can be complemented by the possible linkage of these results to funding trends (awards made by NIH, NSF, society grants, etc) and the comparison of the research presented at professional meetings with these publication trends.

In addition, we hope to predict future directions in animal behavior research. For example we may wish to discover, why more traditional ethological approaches focusing primarily on mechanism and development have been overtaken by the recent popularity of approaches embedded in evolutionary theory, population biology and behavioral ecology? Observations on whether authors studying 'hot topics' publish more are of interest to the domain experts. Further we may want to find the number of articles published by particular authors, their areas of interests, their institutional addresses and gender.

Previous studies have highlighted a potential bias relating to the sex and nationality of authors in the reviewing process of manuscripts in other life sciences [6]. These and other observations dispel significant information about the domain and can be corroborated by the mapping of the knowledge domain.



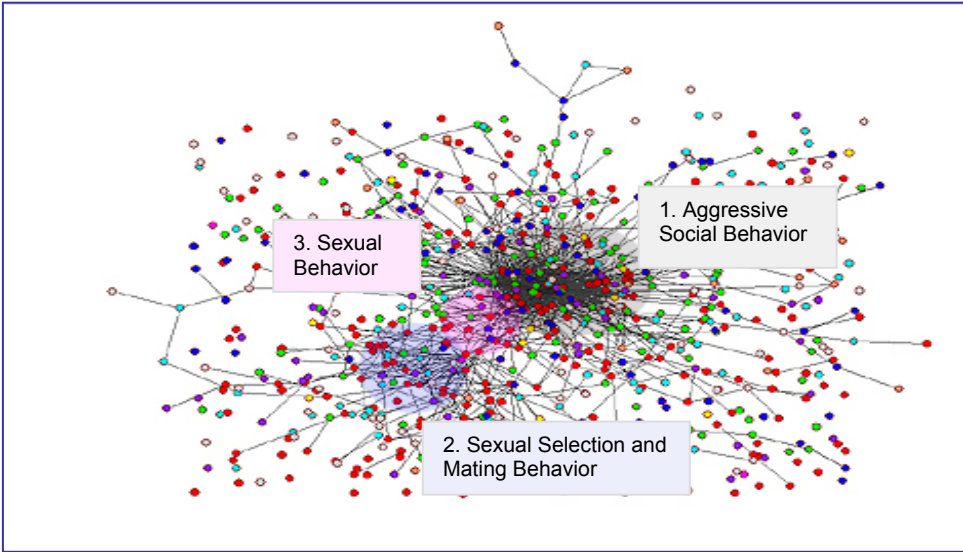


Figure 6:
Semantic Document Space 1997

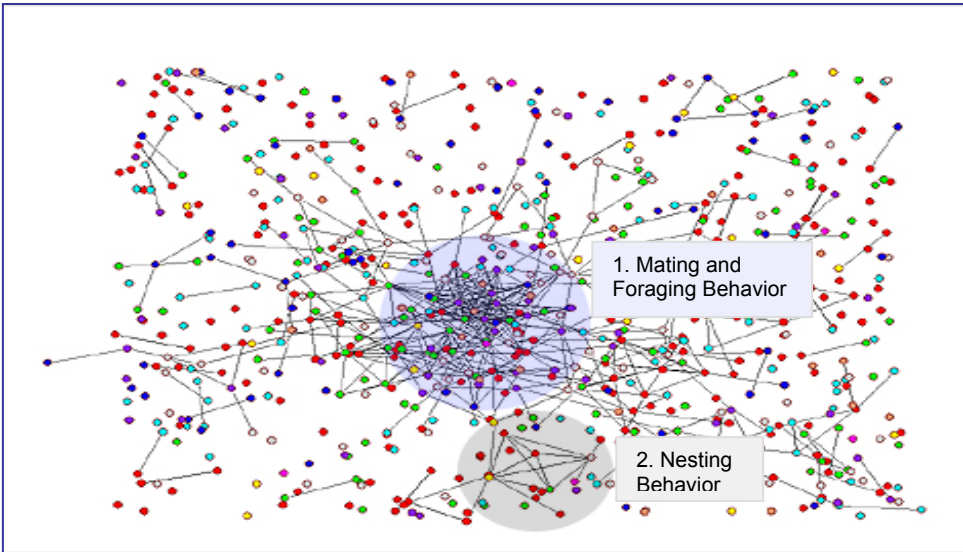


Figure 7:
Semantic Document Space 2000

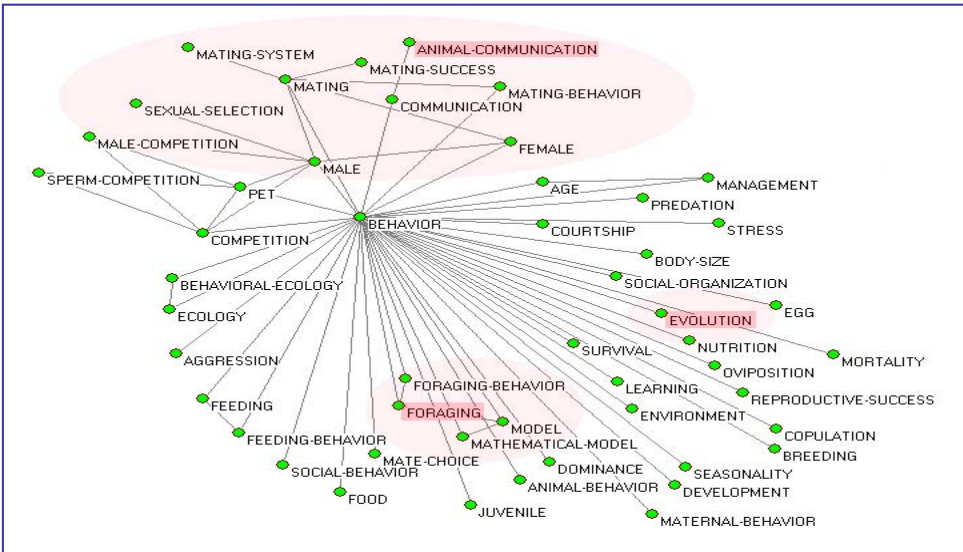


Figure 8:
Keyword space in 1994

