English
Research Express@NCKU Volume 6 Issue 10 - December 12, 2008
[http://research.ncku.edu.tw/re/articles/e/20081212/1.html]

HAL-Based Evolutionary Inference for Pattern Induction From Psychiatry Web Resources

Liang-Chih Yu[1], Chung-Hsien Wu[1,*], Jui-Feng Yeh[2], and Fong-Lin Jang[3]

[1]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C.
[2]Department of Computer Science and Information Engineering, National Chiayi University, Chiayi, Taiwan, R.O.C.
[3]Department of Psychiatry, Chi-Mei Medical Center, Tainan, Taiwan, R.O.C.
chwu@csie.ncku.edu.tw

IEEE Trans. Evolutionary Computation, Vol. 12, No. 2, pp. 160-170, April 2008.

With the increased incidence of depression-related disorders, various Internet-based psychiatric services have emerged for individuals suffering from negative or stressful life events, such as the death of a family member, an argument with a spouse or the loss of a job, along with depressive symptoms, such as suicidal tendencies and anxiety. Individuals under these circumstances often search for help from psychiatric web sites by describing their mental health problems using message boards and other services, thus producing thousands of psychiatric documents, called psychiatry web resources. Knowing these negative life events can therefore enable psychiatric web sites to provide automatic psychiatric services.

Negative life events are often expressed in natural language segments (e.g., sentences, text passages). A critical step in identifying these segments is to transform the natural language segments into machine-interpretable semantic representation. This step involves the extraction of key patterns from the segments. Consider the following example.

*Two years ago, I **lost** my **parents**.* *(Negative life event)*
Since then, I have tried to kill myself several times. *(Suicide)*

In this example, the pattern comprises two words, indicating that the subject suffered from a negative life event triggering the symptom “Suicide”. A pattern can be considered as a semantically plausible combination of *k* words, where *k* denotes the length of the pattern. Accordingly, a pattern has a variable length. This study presents an evolutionary text-mining framework capable of inducing variable-length patterns from *unannotated* psychiatry web resources. The proposed framework, as shown in Fig. 1, can be divided into two parts: 1) a cognitive motivated model such as *Hyperspace Analog to Language (HAL), and 2) an Evolutionary Inference Algorithm (EIA).*

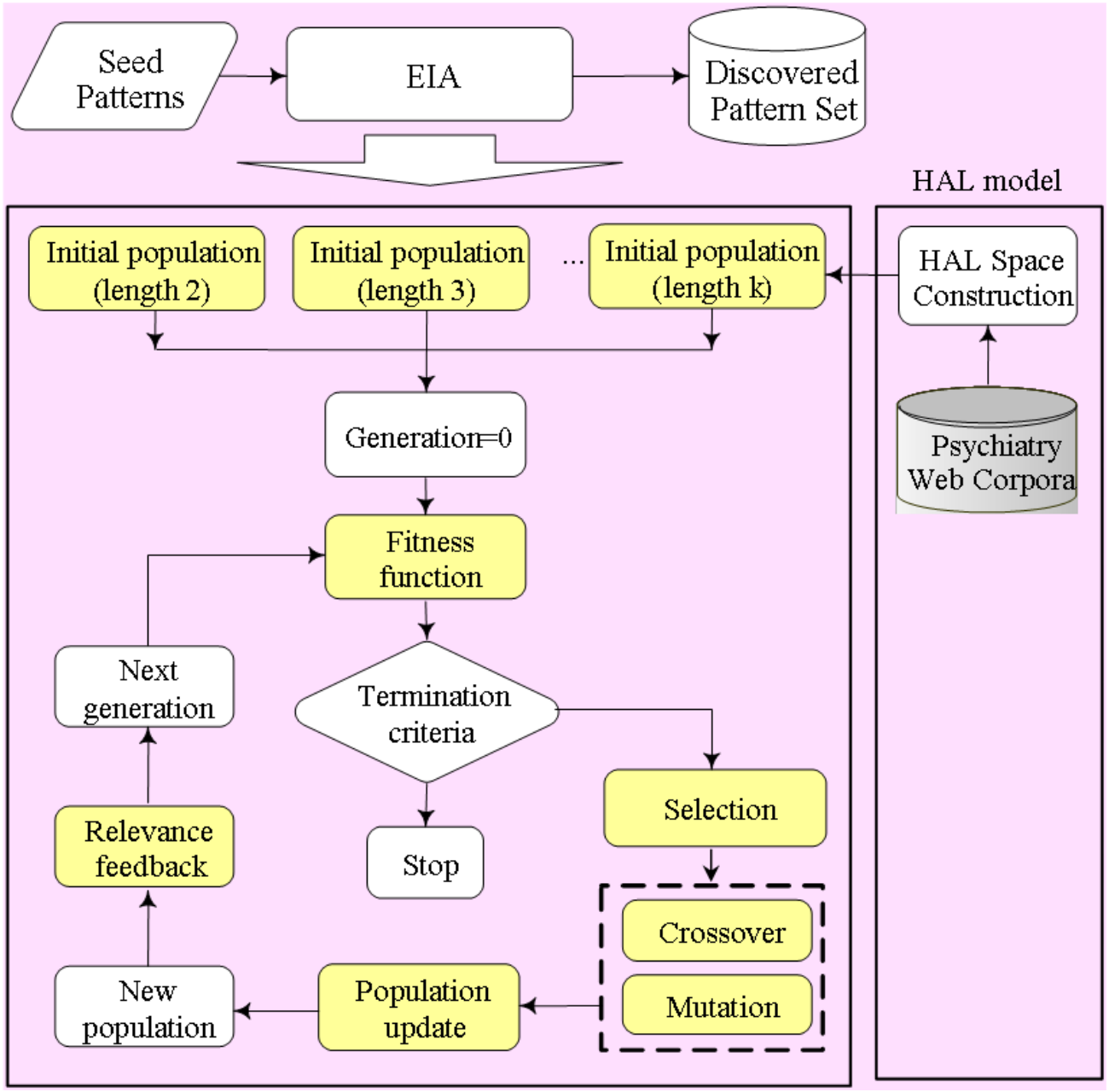

Fig. 1. Framework for variable-length pattern induction.

The HAL model constructs a high-dimensional context space to represent words as well as patterns. Each word/pattern in the HAL space is represented as a vector of its context, which means that the sense of a word/pattern can be inferred from its context. This notion is derived from observations of human behavior. Restated, human beings may determine the sense of an unknown word by referring to the words appearing in its context. Based on the cognitive behavior, two words/patterns sharing more common contexts are more similar semantically. Figure 2 shows an example of the context information of the words “boss”, “chief” and “flower”. “Boss” and “chief” have quite similar contexts, but these are quite different from the context of “flower”. Accordingly, the words “boss” and “chief” are more similar to each other semantically than “boss” and “flower”, and “chief” and “flower”.

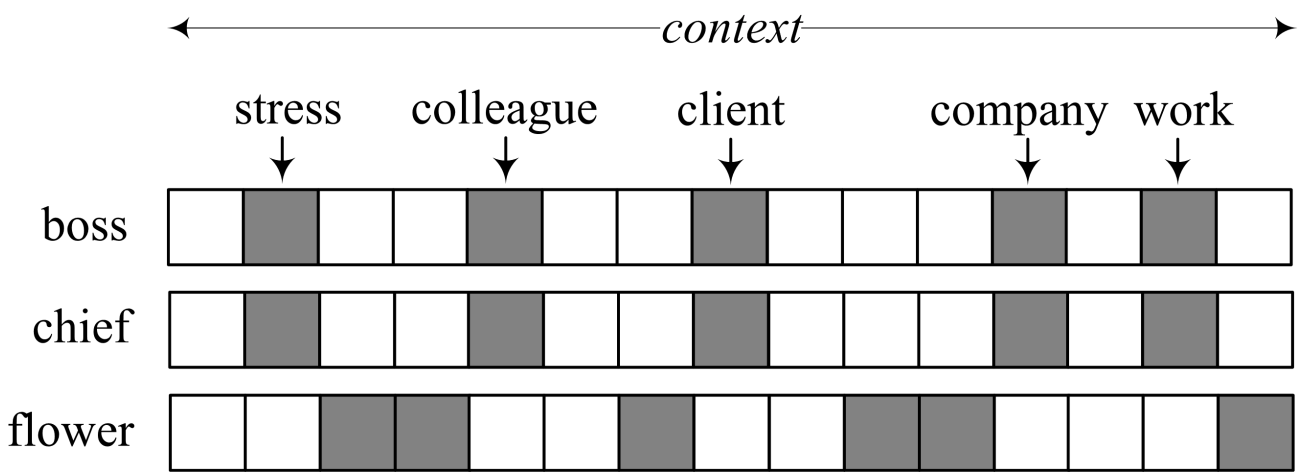

Fig. 2. Example of the HAL model. The dimensions with a gray shadow denote the frequent words in the context of the target word.

The EIA bootstraps with a small set of seed patterns to induce additional relevant patterns once the HAL space is built. The EIA then creates the initial populations of the patterns with different lengths (from 2 to *k*). Each pattern of length *k* is created by selecting *k* distinct words from the vocabulary. Both the patterns and the seed patterns are represented by combining their constituent words over the HAL space. After initializing each population, i.e., generation=0, the *fitness function* is adopted to measure the fitness (similarity) of each pattern in each population based on the context information provided by the HAL model. Once all the populations are measured, the *selection* process chooses a number of patterns to be the parents according to their fitness values. The variation operators, i.e., *crossover* and *mutation*, are then applied to produce the offspring. The offspring is also evaluated by the fitness function, and the superior offspring replaces the inferior parents to form a new population. The *relevance feedback* is then applied to identify the relevant patterns in the population. This information can be adopted to refine the seed pattern to improve its similarity to the relevant set. The refined seed pattern is taken as the reference basis in the next generation. The induction process is performed iteratively until the termination criteria are satisfied. Table 1 shows parts of the seed patterns and the induced patterns with the seed pattern as input.

Table 1. Seed patterns and induced patterns.

Types	Seed Pattern	Pattern Induction from Web
Family	<son, injure>, <husband, argue> →	<husband, fight> <husband, yell> <wife, argue> <spouse, fight> <husband, fight, money > <wife, argue, money>
Love	<marriage, break>, <wife, cheat>	
School	<teacher, blame>, <exam, fail>	
Work	<salary, cut>, < work, stop>	
Social	<friend, die>	

To evaluate the performance of the EIA, a total of 5,000 psychiatric documents were collected from the professional mental health web sites, such as PsychPark (http://www.psychpark.org) and John Tung Foundation (http://www.jtf.org.tw). The baseline system used herein is the association pattern mining

(*Apriori* algorithm), which is a supervised corpus-based approach widely used in data mining community. The EIA and *Apriori* algorithm were then performed, respectively, for pattern induction, and the induced patterns were evaluated on the coverage of real data provided by 15 human subjects. The subjects provided their experienced negative life events in the form of natural language sentences. A total of 69 sentences were gathered as the test set. The evaluation metric adopted in this experiment was the *out-of-pattern (OOP)* rate, a ratio of unseen patterns occurred in the test set, which was calculated as the number of test sentences with a pattern not occurring in the set of discovered patterns, divided by the total number of test sentences. Additionally, a *sign test* for pairwise comparison was adopted to determine whether the performance difference was statistically significant. Table 2 shows the OOP rates of the EIA-based approaches and Apriori algorithm, where the rows denoted by *_*_ *Multiple* represent the OOP rates of the EIA obtained after 30 experiments, and the rows denoted by *_*_ *Worst* represent the worst OOP rates over the 30 experiments.

Table 2. Comparison of OOP rates between EIA and Apriori algorithm.

	OOP Rate	Reduction (%)			
		over Apriori	over EIA_Worst	over EIA_Multiple	over EIA_RF_Worst
Apriori	60.9% (42/69)	—	—	—	—
EIA_Worst	97.1% (67/69)	+59.4	—	—	—
EIA_Multiple	42.0% (29/69)	-31.0#	-56.7#	—	—
EIA_RF_Worst	40.6% (28/69)	-33.3#	-58.2#	-3.3	—
EIA_RF_Multiple	27.5% (19/69)	-54.8#	-71.7#	-34.5#	-32.3#

performance difference is statistically significant ($p<0.05$)

The experimental results indicate that the EIA is more effective than *Apriori* algorithm mainly due to the incorporation of the HAL model, the use of relevance feedback, and multiple experiments. The HAL model provides an informative infrastructure to represent the patterns in a high-dimensional context space. Based on the HAL space, the EIA can bootstrap with a set of seed patterns, and then induce more relevant patterns from unannotated corpora with the help of relevance feedback. Additionally, the use of EIA multiple experiments also improves performance. The proposed evolutionary framework only needs a small amount of annotated data, thus reducing the reliance on the availability of large annotated corpora, as required by corpus-based approaches.