The Smell Network

Ritesh Kumar\textsuperscript{1,2(✉)}, Rishemjit Kaur\textsuperscript{3}, and Amol P. Bhondekar\textsuperscript{1,2}

\textsuperscript{1} CSIR-Central Scientific Instruments Organisation, Chandigarh, India
riteshkr@csio.res.in
\textsuperscript{2} Academy of Scientific and Innovative Research, New Delhi, India
\textsuperscript{3} Nagoya University, Nagoya, Japan

Abstract. The smell of a molecule is subjective, because there is a variability in its representative language. The reporting is done according to the vocabulary repertoire of human subjects and researchers concerned. The olfactory databases thus consist of molecules and their smell characteristics defined by words. In this paper, we have demonstrated a network based approach based on the words to understand the perceptual universe. We defined perceptual communities based on the normalized co-occurrence network and hence propose the perceptual classes. We find the characteristics of this perceptual social network. We have also proposed a generative LDA-based topic modeling approach for topic detection in olfactory databases. This is for the first time that an objective approach to defining perceptual classes has been carried out which confirms with many subjective analyses that has been done till now. This work may open new avenues towards understanding the relationship between language and olfaction besides objectively defining perceptual classes.

Keywords: Social network analysis · Clustering coefficient · Community detection · Latent diritchlet allocation · Olfaction · Data mining

1 Introduction

Recently, network analysis has grabbed much attention due to its clear representation in terms of entities and relationship and, they have provided some really interesting insights into the data they represent [1]. The present work aims at discovering inherent statistical structure in large chemical and perceptual databases available online in order to derive principles for predicting odor perception from the chemical structure of odorants. An adjacency list of odorants on the basis of perceived smell was created. Further, a smell network was created in which each odorant forms the node and weight of edge between them shows the normalized number of smells they share with each other. The results provide useful insights into the odor space such as the special characteristics of the smell network.

The last decade has seen a tremendous surge in natural language processing based software and their application to different areas. The mathematical techniques behind these works have given an impetus to even the social sciences and literature[2–4]. The most famous works have been to automatically detect topics and classify large corpus
of documents into well organized corpus. Probabilistic topic modeling techniques provide comprehensive algorithms to define topics from which we can make a definite sense [5]. The techniques have been used on many kinds of data and have found application in finding patterns in genetic data, images and social networks etc. [6,7,8]. The last decade has seen a tremendous surge in natural language processing based software and their application to different areas. The mathematical techniques behind these works have given an impetus to even the social sciences and literature [2–4].

The most famous works have been to automatically detect topics and classify large corpus of documents into well organized corpus. Probabilistic topic modeling techniques provide comprehensive algorithms to define topics from which we can make a definite sense[5]. The techniques have been used on many kinds of data and have found application in finding patterns in genetic data, images and social networks etc. [6,7,8].

Olfaction has been a very subjective and contentious of our senses owing to large variations in its reporting which is partly due to the fact that people experience it very differently [9,10,11]. The only non-invasive method of its reporting in humans is via language. Here, our language abilities have either failed us or we have not dug enough [12]. There is no standard way of representation of the smell of an odour molecule. The words spoken by the subjects vary too much. Hence, it is imperative to understand the theme behind the words spoken by the individuals. There are a lot of potential benefits of such kind of analysis. First, it will let us understand the relationship between language and olfaction. Second, it will let us rationalize the enigma behind the smell representation. Third, it will give an objectivity to the olfactory perceptual representation, which we can be further utilised in designing odour molecules. It can also help in designing searchable databases having meaningful smell topics and their relationship with each other.

1.1 Database Description

The present study looks into 4 different databases available publically i.e. SuperScent [13], Leon and Johnson [14] (LJ), Flavornet [15] and GoodScents [16]. The prime feature of all these databases is odor molecules described by some words or percepts and most of the times their molecular references. It should be noted that the variance of perceptual descriptors in these databases cannot be fully established due to the lack of information related to the odor experiments for them. Some discrepancies in the molecular reference entries were validated and removed by comparing the given molecular references entries and the calculated molecular weights by software E-DRAGON(“Molecular Descriptors for Chemoinformatics. 2 Edn. Wiley-VCH. 2009). Then, some curation was done on these databases for creation of the feature sets. Firstly, the words describing the perceptual qualities of the molecules were tokenized to result in a set of perceptual descriptors (eg: for acetal perceptual qualities were specified as “ether green nut earthy sweet vegetable” these were tokenized as “ether” “green” “nut” “earthy” “sweet” “vegetable”). Overall, from the tokens for all the molecules, conjunctions (e.g. “and”), adverbs (e.g. “less”, “somewhat”), suffixes (e.g. “like”, “note”), auxiliary verbs (e.g. “has”) and some other words which dont convey