Concept Learning and Information Inferencing on a High Dimensional Semantic Space

Dawei Song and Peter Bruza
CRC for Enterprise Distributed Systems Technology
The University of Queensland
QLD 4072 Australia
{dsong, bruza}@dstc.edu.au

Richard Cole
School of Information Technology and Electrical Engineering
The University of Queensland
QLD 4072 Australia
rcole@itee.uq.edu.au

Abstract

How to automatically capture a significant portion of relevant background knowledge and keep it up-to-date has been a challenging problem encountered in current research on logic based information retrieval. This paper addresses this problem by investigating various information inference mechanisms based on a high dimensional semantic space constructed from a text corpus using the Hyperspace Analogue to Language (HAL) model. Additionally, the Singular Value Decomposition (SVD) algorithm is considered as an alternative way to enhance the quality of the HAL matrix as well as a mechanism of infering implicit associations. The different characteristics of these inference mechanisms are demonstrated using examples from the Reuters-21578 collection. Our hope is that the techniques discussed in this paper provide a basis for logic based IR to progress to large scale applications.

Keywords: Logic-based Information Retrieval, Information Inference

1 Introduction

Logic-based information retrieval views retrieval as a process of plausibly inferring the query from the document. Since Van Rijsbergen proposed the logical uncertainty principle in one of the area's founding papers [9], a number of logical IR models have been developed. These logical frameworks exist in the realm of symbolic processing. This realm is characterized by tokens representing atomic propositions. Atomic propositions can be composed into more complete propositions via connectives. Inference is a sequential process proceeding from assumptions to a conclusion by applying rules of inference. Despite their advantage in modelling information transformation and their expressive power, these logical frameworks have difficulty when applied in large scale applications. We believe that the central challenge involved in facilitating effective inference in these frameworks is the construction and maintenance of a knowledge base in forms of logical implications (propositional logic), defaults (default theory), constraints (situation theory), preferential entailments (preference logic), background theory (abduction), etc.

The background knowledge could be generally termed as various "aboutness" relationships, which are sensitive to contexts. For example, "penguin" is normally about "bird"; "penguin" uttered the context of "books, UK", on the other hand", is about "publisher". Human can generally make robust judgments about what information fragments are, or are not about, even when the fragments are brief or incomplete. The process of automatically making such "aboutness" judgments has been referred to as informational inference in our recent work [8].

It is an enormous specification task to capture a significant portion of relevant background knowledge and keep it up-to-date. Traditionally this has been done by either manually collecting domain expert's knowledge, or by using some sort of pre-built thesaurus. The knowledge base constructed in this way is static and cannot reflect our ever changing cognitive environment. Moreover, inference generally cannot be performed without an explicit attempt to representation and capture context. Another barrier to the construction of large scale systems is the computational complexity inherent in symbolic inference.

We have recently developed an information flow inference model to automatically discover how strongly a concept Y is informationally contained within another concept X. Words and concepts are represented as vectors in a high dimensional semantic space constructed from a text corpus. Information flow computation between vectors is proposed as a means of suggesting potentially interesting implicit associations between concepts. The information flow model has been successfully applied to facilitate the contextual knowledge, represented as logical implications, in a belief revision framework for adaptive information retrieval [5]. In this work, a retrieval context refers to an information seeker's background, long term search goals, etc. The information flow model is used to automatically derive the initial retrieval context from a query. When the user's information need changes, the retrieval context is then updated by a computerised implementation of the AGM belief revision system. To our best knowledge, this is the first logic based IR model which has been applied and evaluated on a large collection (AP).

Nevertheless, the information flow is obviously not the only way to draw information inference. The purpose of this paper is to address a number of information inference mechanisms from a high dimensional semantic space and show their different natures.

2 Semantic spaces

A human encountering a new concept derives its meaning via an accumulation of experience of the contexts in which the concept appears. The meaning of a word is captured by examining its co-occurrence patterns with other words in the language use (e.g., a corpus of texts). There have been two major classes of semantic space models: document spaces and word spaces. The former represents words as vector spaces of text fragments (e.g. documents, paragraphs, etc) in which they appear. A notable example is the Latent Semantic Analysis (LSA). The latter represents words as vector spaces of other words, which occur with the target words within a certain distance (e.g., a window size). The weighting scheme used in the Hyper-space Analogue to Language (HAL) model makes term term association inversely proportional to the distance between the context and target words. The semantic spaces are always high dimensional, e.g., Burgess et al. constructed a 70,000x70,000 HAL vector spaces from a 300 million word text collection gathered from Usenet [2], while Landauer et al. generated a 30,000 dimensional document vector for each of 60,000 words from an encyclopaedia [4]. Such huge dimensionality is computationally expensive. A dimensional reduction is always applied prior to further processing. For example, a Singular Value Decomposition (SVD) is applied in LSA [4] to extract roughly the 300 most important dimensions. It also performs some sort of induction, for example some words not occurring in a passage are inferred and the weights of some originally occurring terms are changed.

The semantic space models have demonstrated cognitive compatibility with human processing. For example, Burgess and Lund showed that HAL vectors can be used to simulate semantic, grammatical and abstract categorizations. The synonymy problem can be simulated nicely by semantic similarities, which fundamentally capture the degree substitutability between concepts. As an example, Landauer et al. [4] evaluated semantic similarity computed from the LSA matrix using the vocabulary test (synonyms) data in TOEFL and got 52.7% correct responses in the best case (300-325 dimensions). Meanwhile, Burgess et al. [2] also showed via cognitive experiments that "human participants were able to use the context neighbourhoods that HAL generates to match words with similar items and to derive the word (or a similar word) from the neighbourhood, thus demonstrating the cognitive compatibility of the representations with human processing".

For the purpose of this paper, we choose HAL model to represent information. Meanwhile, Singular Value Decomposition (SVD), which plays a fundamental role in LSA, is considered as a kind of inference mechanism via dimensional reduction on the HAL space.

	the	effects	of	spreading	Pollution	on	Population	Atlantic	salmon
the		1	2	3	4	5			
effects	5								
of	8	5		1	2	3	5		
spreading	3	4	5						
pollution	2	3	4	5					
on	1	2	3	4	5				
population	5		1	2	3	4			
atlantic	3		5		1	2	4		
salmon	2		4			1	3	5	

Table 1: Example of a HAL space

3 Concept Formation via Hyperspace Analogue to Language

Given an n-word vocabulary, the HAL space is an $n \times n$ matrix constructed by moving a window of length l over the corpus by one word increments. Given two words w_1 and w_2, whose distance within the window is d, the weight of association between them is computed by $l - d + 1$. After traversing the corpus, an accumulated co-occurrence matrix for all the words in a target vocabulary is produced. HAL is direction sensitive: the co-occurrence information for words preceding each word and co-occurrence information for words following each word are recorded separately by row and column vectors. By way of illustration, the HAL space for the example text "The effects of spreading pollution on the population of Atlantic salmon" is depicted in Table 1. As an illustration, the term "effects" appears after the term "spreading" in the window and their distance is two words. The value of the cell corresponding to (spreading, effects) can then be computed as: $5 - 2 + 1 = 4$.

Table 1 shows how the row vectors encode preceding word order and the column vectors encode posterior word order. For the purposes of this paper, it unnecessary to preserve order information, so the HAL vector of a word is represented by the addition of its row and column vectors. The quality of HAL vectors is influenced by the window size; the longer the window, the higher the chance of representing spurious associations between terms. Window sizes of eight and ten have been used in various studies [2, 1].

According to the Zipf's law (power scaling between a word's occurrence frequency and its ranking by frequency), a word's power for discriminating content relies on it being neither too rare nor too common. During the pre-processing of a text corpus, extremely infrequent words can be removed by setting a frequency threshold, say 5. Extremely frequent words are collected as stop words, e.g., "the", "of", etc. Stop words do not carry much semantically useful information and can be removed by using a stop word list.

Even after getting rid of the rare and stop words, however, the weighting scheme of HAL is still frequency biased — a small number of most frequent words tend to get higher weights in any HAL vector, due to their high frequency and so caused chance co-occurrence. The high frequent words may not be the most informative ones. For example, "corp" in IBM vector is highly weighted, indicating that it can be used to describe the vector IBM. However, it also appears in many other vectors from a corpus of financial news, so that it may not be very useful in discriminating these vectors. On the other hand, less frequent terms may not necessarily be less important.

Motivated by experience with Inverse Document Frequency (IDF) in IR community, we propose the inverse vector frequency (IVF) be used to measure the informativeness of a word.

$$\mathrm{IVF}(w) = \frac{\log\left(\frac{N+0.5}{n}\right)}{\log\left(N+1\right)} \tag{1}$$

where N is the total number of HAL vectors (i.e., number of unique words), and n is the number of dimensions of the HAL vector for word w (i.e., the number of vectors with w as a dimension).

This formula is actually the Inquery's IDF formula. The more vectors w appears in, the lower its IVF value. The IVF function produces values in the range $(0, 1)$.

	Number of dims	collection frequency	IVF	Entropy
mln	10134	26732	0.043	6.024
dlrs	10045	21266	0.044	6.397
pct	9795	18045	0.047	6.961
company	9707	9697	0.048	7.531
market	7831	5956	0.070	7.566
corp	7513	7157	0.074	7.332
billion	6820	10726	0.085	6.256
president	5903	2741	0.10	7.353
administration	3232	829	0.162	7.127
reagan	3290	1272	0.160	6.960
north	2702	648	0.181	7.026
arms	1106	226	0.211	6.099
scandal	825	125	0.304	6.012
rebels	423	68	0.373	5.366
contra	393	62	0.381	5.174
tower	374	45	0.386	5.443
poindexter	281	57	0.415	5.079
nicaraguan	266	36	0.421	5.019
oliver	191	24	0.455	4.671

Table 2: Example of IVF and Entropy values of words

Alternatively, the entropy method could be applied to HAL vectors. This method has been used by Landauer et al. [4] for LSA. The equation for entropy is:

$$\text{Entropy}_{\text{HAL}}(w) = -\sum_{i=1}^{n} \Pr(t_i|w) \log \Pr(t_i|w) \tag{2}$$

where n is the number of dimensions in the HAL vector for w, denoted $\text{HAL}(w)$, and t_i is the i-th dimension in $\text{HAL}(w)$, denoted $\text{HAL}(t_i|w)$.

$$\Pr(t_i|w) = \frac{\text{HAL}(t_i|w)}{\sum_{j=1}^{n} \text{HAL}(t_j|w)} \tag{3}$$

Table 2 lists some words from the Reuters collection, wherein the total number of words is 15,547 after removing stop words and the least frequent words (< 5).

It seems that the number of dimensions in the HAL vector of a word (i.e., how many HAL vectors in which the word appears as a dimension) is biased to the word's collection frequency; as a consequence, the high frequent words tend to have lower IVF values. On the other hand, the entropy has to do with the distribution of words. The more evenly distributed a word is over the HAL vectors, the larger its entropy value is. The entropy value of a word does not necessarily seem to be proportionally related to the frequency, as the IVF does.

Related to this issue, a column variance method was used by Burgess et al. [2]. Lowe [6] has shown that this method is still high frequency biased. Instead, he proposed a log-odds-ratio method to factor out chance co-occurrence.

In this paper, we just use the IVF method to reweight HAL vectors. Each dimension in a HAL vector can be re-weighted by multiplying its HAL weight and its IVF value. We will leave the investigation and comparison of other kinds of smoothing techniques as future work.

The following example shows how the weights and rankings of "billion" and "scandal" in the "reagan" vector change after applying IVF. It is desirable that the ranking of non-informative dimension "billion" decreases while

the ranking of "scandal" increases.

$$\begin{aligned}
\mathrm{HAL}(billion|reagan) &= 416 \overset{\text{normalize}}{\longrightarrow} 0.08 \overset{\text{rank}}{\longrightarrow} (15) \\
\mathrm{HAL}_{ivf}(billion|reagan) &= 35.2 \overset{\text{normalize}}{\longrightarrow} 0.05 \overset{\text{rank}}{\longrightarrow} (55) \\
\mathrm{HAL}(scandal|reagan) &= 145 \overset{\text{normalize}}{\longrightarrow} 0.03 \overset{\text{rank}}{\longrightarrow} (75) \\
\mathrm{HAL}_{ivf}(scandal|reagan) &= 44.1 \overset{\text{normalize}}{\longrightarrow} 0.05 \overset{\text{rank}}{\longrightarrow} (33)
\end{aligned}$$

In summary, by combining HAL, IVF values, removing stop words and the least frequent terms, we can build a more informative semantic space.

More formally, a concept c is a vector representation: $c = \langle p_1, \ldots, p_n \rangle$ where $p_1, \ldots, p_n$ are called dimensions of c, n is the dimensionality of the HAL space, and w_{cp_i} denotes the weight of p_i in the vector representation of c. A dimension is termed a property if its weight is greater than zero. A property p_i of a concept is a termed a *quality property* iff $w_{cp_i} > \delta$, where $delta$ is a non-zero threshold value. Let $\mathrm{QP}_\delta(c)$ denote the set of quality properties of concept c. $\mathrm{QP}_\mu(c)$ will be used to denote the set of quality properties above mean value, and $\mathrm{QP}(c)$ is short for $\mathrm{QP}_0(c)$. We have proposed to use the following normalization algorithm for a dimension p_j in concept c_i [7].

$$w_{c_i p_j} = \frac{w_{c_i p_j}}{\sqrt{\sum_k w_{c_i p_k}^2}} \tag{4}$$

The following example is the HAL vector for concept "Reagan" from the Reuters collection, which is related to a number of different contexts, e.g., the general role of Reagan as the president of US; the Iran-Contra scandal; and the US and Japan trade war; etc.

---

reagan = ⟨president: 0.46, administration: 0.42, veto: 0.20, reagan: 0.19, house: 0.16, congress: 0.14, white: 0.13, budget: 0.13, bill: 0.12, senate: 0.11, iran: 0.11, arms: 0.11, tariffs: 0.11, trade: 0.10, ronald: 0.09, nakasone: 0.08, tax: 0.08, baker: 0.07, sanctions: 0.07, japan: 0.07, highway: 0.07, impose: 0.06, retaliate: 0.06, fitzwater: 0.06, summit: 0.06, legislation: 0.06, volcker: 0.06, decision: 0.06, japanese: 0.06, vetoed: 0.05, republicans: 0.05, secretary: 0.05, scandal: 0.05, televised: 0.05, contra: 0.05, washington: 0.05, democrats: 0.05, congressional: 0.05, officials: 0.05, economic: 0.05, tower: 0.05, republican: 0.05, acid: 0.05, taxes: 0.05, policy: 0.05, talks: 0.05, poindexter: 0.05, opposes: 0.05, speech: 0.04, imposed: 0.04, aide: 0.04, staff: 0.04, miller: 0.04, dole: 0.04, billion: 0.04, deficit: 0.04, retaliation: 0.04, override: 0.04, action: 0.04, advisers: 0.04, contras: 0.04, gulf: 0.04, opposed: 0.04, ...⟩

---

4 Concept Learning via Concept Combination

Our ability to combine concepts and, in particular, to understand new combinations of concepts is a remarkable feature of human thinking. Regarding to the context of this paper, combinations of words in document title may represent a single underlying concept, for example, "Reagan and scandal". An important intuition is that one concept can dominate the other within a combination. For example, the term "scandal" can be considered to dominate the term "reagan" because it serves as a context of "Reagan" and in turn carries more of the information in the combination.

Song and Bruza [7] proposed a concept combination heuristic, which is essentially a restricted form of vector addition whereby quality properties shared by both concepts are emphasized, the weights of the properties in the dominant concept are re-scaled higher, and the resulting vector from the combination heuristic is normalized to smooth out variations due to differing number of contexts the respective concepts appear in (Refer to [7] for details of the heuristic).

In order to deploy the information flow model in an experimental setting, dominance is determined by multiplying the term frequency (tf) by the inverse document frequency (idf) value of the term. More specifically, terms can be re-ranked according to tf$*$idf. Assume such a ranking of terms: $t_1, \ldots, t_m$, $m > 1$. Terms t_1 and t_2 can be combined using the concept combination heuristic described above resulting in the combined concept $t_1 \oplus t_2$, whereby t_1 dominates t_2 (as it is higher in the ranking). For this combined concept, its degree of dominance is the average of the respective tf$*$idf scores of t_1 and t_2. The process recurses down the ranking resulting in the composed "concept" $((\ldots(t_1 \oplus t_2) \oplus t_3) \oplus \cdots \oplus t_m)$. If there is a single term ($m = 1$), it's corresponding normalized HAL vector is used as the combination vector.

Given two concepts $c_1 = \langle w_{c_1 p_1}, \ldots, w_{c_1 p_n} \rangle$ and $c_2 = \langle w_{c_2 p_1}, \ldots, w_{c_2 p_n} \rangle$. The resulting combined concept is denoted $c_1 \oplus c_2$. The following is a fragment of the vector resulting from the combination of the HAL vectors for "Scandal" and "Reagan":

---

scandal = ⟨arms: 0.57, iran: 0.47, contra: 0.22, insider: 0.16, vw: 0.15, reagan: 0.13, currency: 0.12, volkswagen: 0.12, investigating: 0.11, probing: 0.11, prosecutor: 0.11, spy: 0.10, poindexter: 0.09, vowgf: 0.09, sex: 0.09, immunity: 0.08, arrested: 0.08, ...⟩

---

scandal $\oplus$ **reagan** = ⟨ arms: 0.56, iran: 0.47, contra: 0.22, reagan: 0.17, administration: 0.17, president: 0.16, insider: 0.14, investigating: 0.11, currency: 0.11, probing: 0.10, prosecutor: 0.10, poindexter: 0.10, senate: 0.10, spy: 0.10, house: 0.08, sex: 0.08, immunity: 0.08, wall: 0.07, tower: 0.07, trading: 0.07, veto: 0.07, street: 0.06, congress: 0.06, investigation: 0.06, security: 0.06, wake: 0.05, nsc: 0.05, embassy: 0.05, affair: 0.05, broke: 0.05, white: 0.05, investigators: 0.05, indictments: 0.05, moscow: 0.05, shultz: 0.05, grant: 0.05, cia: 0.05, page: 0.05, special: 0.05, witness: 0.05, bill: 0.04, resigned: 0.04, budget: 0.04, rebels: 0.04, diversion: 0.04, november: 0.04, televised: 0.04, baker: 0.04, allegedly: 0.04, contras: 0.04, profits: 0.04, congressional: 0.04, wallison: 0.04, diverted: 0.04, nicaraguan: 0.04, misled: 0.04, key: 0.04, felony: 0.04, speech: 0.04, role: 0.04, ...⟩

We can observe that "Scandal" and "Reagan" have been conditioned by each other, indicating the Iran-contra scandal involving President Reagan.

Note that we use individual words as dimensions in a semantic space. The phrases, e.g., "iran contra scandal", could be mimicked by concept combination heuristic. This gives a more general and flexibile way to create "new concept" and derive its meaning from any arbitrary composition of certain related terms, not being limited to syntatically valid phrases.

5 Information Inference on the HAL Space

5.1 Inference via Defaults

The extremely highly weighted dimensions in a HAL vector can be considered as defaults of the corresponding concept. This is a kind of context free inference, indicating the concept refers to these defaults in any context (under the scope of the collection).

Definition 1 (Defaults) *Given a concept c, a dimension p_i in c's HAL vector is a default of c (denoted $c \sim p_i$) if and only if*

$$w_{cp_i} \geq l * \mathrm{STD}(c) + \mathrm{MEAN}(c)$$

where $\mathrm{STD}(c)$ is the standard deviation of weights of all the non-zero dimentions in the HAL vector of c, $\mathrm{MEAN}(c)$ is the mean weights of all the non-zero dimentions in the HAL vector of c and $l \geq 0$ is a parameter to be determined experimentally. It is normally a very large number. In our experiments we set it to be 13.

Examples:

Reagan $\sim$ president;

Reagan $\sim$ administration;

5.2 Inference via Lexical Associations

The highly weighted dimensions in the HAL vector of a concept c are those often lexically co-occurring with c in the collection. These dimensions are termed as lexical associations of c. Note that all defaults are obviously lexical associations, but not vice versa. For the ease of use, we may exclude the defaults from the definition of lexical associations.

Definition 2 (Lexical Associations) *Given a concept c, a dimension p_i in c's HAL vector is a lexical association of c (denoted $c \gg p_i$) if and only if*

$$w_{cp_i} \geq \theta \text{ and } p_i \notin \{p \mid c \sim p\}$$

where θ is a threshold value or function. An example of such a threshold function may be selecting the dimensions whose weights are above the mean of the weights of all the dimensions in the vector.

Examples:

Reagan $\gg$ congress

Reagan $\gg$ white, house

...

5.3 Inference via Overlapping Semantic Associations

The concepts, which occur in the similar contexts, tend to be similar to each other in meaning. For example, nurse and doctor are similar in semantics to each other, as they always experience the same contexts, i.e., hospital, patients, etc. The similarity can be measured by the angle (Cosine) or Euclidean distance between two word vectors in the semantic space. These functions (shown below) are symmetric and essentially measure the degree of overlapping between two vectors.

$$\text{Similarity-cosine}(c_i, c_j) = \frac{\sum_k w_{c_i p_k} w_{c_j p_k}}{\sqrt{\left(\sum_k w_{c_i p_k}^2\right)\left(\sum_k w_{c_j p_k}^2\right)}} \tag{5}$$

$$\text{Minkowski}(c_i, c_j) = \left(\sum_k \left| w_{c_i p_k} - w_{c_j p_k} \right|^l\right)^{\frac{1}{l}} \tag{6}$$

$$\text{Similarity-minkowski}(c_i, c_j) = e^{-k * \text{Minkowski}(c_i, c_j)} \tag{7}$$

In our experiments, l is set to be 2 (Burgess et al. [2] also used $l = 2$ in their experiments), and k is set to be $1/1500$.

Definition 3 (Semantic Associations) *Given a concept c, another concept c_i is in a semantic association of c (denoted $c \mathrel{|\!\sim} c_i$) if and only if*

$$\text{Similarity}(c, c_j) \geq \lambda$$

where λ is a threshold.

The example below shows the semantic associations of "Reagan" via the Cosine similarity function[1]. Each word is followed by its ranking (in brackets) and the degree of similarity to "Reagan":

[1]The Minkowski function gives similar results

Reagan $\mid\sim$ (Cosine) $\langle$ reagan (1): 0.98 vice (2): 0.67 congress (3): 0.63 ronald (4): 0.62 reagans (5): 0.59 veto (6): 0.58 kenneth (7): 0.58 miller (8): 0.57 suharto (9): 0.57 francesco (10): 0.53 retaliate (11): 0.53 appointed (12): 0.53 aquino (13): 0.52 jose (14): 0.52 resigns (15): 0.51 opposed (16): 0.51 vows (17): 0.50 corazon (18): 0.50 gerald (19): 0.50 opposes (20): 0.49 fusco (21): 0.49 named (22): 0.49 sarney (23): 0.49 legislation (24): 0.48 congressional (25): 0.48 served (26): 0.47 bill (27): 0.46 administration (28): 0.46 urged (29): 0.46 mitterrand (30): 0.46 yoweri (31): 0.45 gemayel (32): 0.45 vetoed (33): 0.45 fidel (34): 0.45 tough (35): 0.45 ferdinand (36): 0.44 jimmy (37): 0.44 urging (38): 0.44 garcia (39): 0.44 jean (40): 0.44 thomas (41): 0.44 robert (42): 0.44 congressmen (43): 0.44 marc (44): 0.44 saddam (45): 0.44 retaliating (46): 0.44 michael (47): 0.44 house (48): 0.44 impose (49): 0.43 passed (50): 0.43 ...$\rangle$

It can be observed that most semantic associations of "Reagan" are those politicians/presidents, for example, "Suharto".

5.4 Inference via Inclusion: Information Flow

Different from the semantic similarity, the intuition of information flow inference (denoted as $\vdash$) is to compute a degree of inclusion between the source and target vectors. Inclusion is a relation over HAL vectors.

Definition 4 (HAL-based information flow)

$$i_1, \ldots, i_k \vdash j \quad \text{iff} \quad \text{degree}(\oplus c_i \lhd c_j) > \lambda$$

where c_i denotes the conceptual representation of token i, and λ is a threshold value. ($\oplus c_i$ refers to the combination of the HAL vectors into a single vector representation representing the combined concept. Details of a concept combination heuristic can be found be in [7]. The degree of inclusion is computed in terms of the ratio of intersecting quality properties of c_i and c_j to the number of quality properties in the source c_i:

$$degree(c_i \lhd c_j) = \frac{\sum\limits_{p_1 \in \text{QP}_\mu(c_i) \cap \text{QP}_\delta(c_j)} w_{c_i p_1}}{\sum\limits_{p_k \in \text{QP}_\mu(c_i)} w_{c_i p_1}}$$

The underlying idea of this definition is to make sure that a majority of the most important quality properties of c_i appear in c_j.

The information flow inference is also biased to the highly frequent targets. High-freq terms, e.g., billion, tend to co-occur with most of other terms. This implies most of other terms are dimensions in the "billion" vector. As a result, almost all the word vectors would have a high degree of inclusion by the "billion" vector. It is interesting to draw a parallel to the tautology in propositional logic. A tautology is a proposition which is always true. It therefore carries no information and can be implied by anything, i.e. $\vdash 1$.

In order to overcome this problem, we consider using IVF values to re-rank the resultant information flows:

$$i_1, \ldots, i_k \vdash \text{IVF}(c_j) * \text{degree}(\oplus c_i, \lhd c_j) > \lambda \tag{8}$$

Example: Information flows (with rankings and degrees) from Scandal $\oplus$ Reagan

Scandal $\oplus$ Reagan $\vdash \langle$ scandal (1): 0.30 poindexter (2): 0.25 contra (3): 0.25 arms (4): 0.23 diversion (5): 0.22 investigators (6): 0.22 affair (7): 0.22 contras (8): 0.22 immunity (9): 0.22 rebels (10): 0.22 testify (11): 0.21 televised (12): 0.21 regan (13): 0.21 nicaraguan (14): 0.20 secret (15): 0.20 probing (16): 0.20 oliver (17): 0.20 tower (18): 0.20 prosecutor (19): 0.20 nsc (20): 0.19 adviser (21): 0.19 resigned (22): 0.19 iran (23): 0.19 aide (24): 0.19 prosecution (25): 0.19 diverted (26): 0.18 republican (27): 0.18 nomination (28): 0.18 speakes (29):

0.18 knew (30): 0.18 senators (31): 0.18 testimony (32): 0.18 investigating (33): 0.18 senator (34): 0.18 byrd (35): 0.18 senate (36): 0.18 illegal (37): 0.18 white (38): 0.18 fired (39): 0.17 cia (40): 0.17 weinberger (41): 0.17 committees (42): 0.17 indictments (43): 0.17 nicaragua (44): 0.17 gates (45): 0.17 questions (46): 0.17 walsh (47): 0.17 millions (48): 0.17 shultz (49): 0.16 constitutionality (50): 0.16 col (51): 0.16 affairs (52): 0.16 bush (53): 0.16 staff (54): 0.16 investigation (55): 0.16 congressional (56): 0.16 george (57): 0.16 howard (58): 0.16 investigate (59): 0.16 criticism (60): 0.16 reporter (61): 0.16 message (62): 0.16 admitted (63): 0.16 caspar (64): 0.16 occasions (65): 0.16 reagan (66): 0.16 questioned (67): 0.16 magazine (68): 0.16 donald (69): 0.16 address (70): 0.16 conversations (71): 0.16 presidential (72): 0.16 defeat (73): 0.16 denial (74): 0.16 opposed (75): 0.16 truth (76): 0.16 refused (77): 0.16 involvement (78): 0.16 wallison (79): 0.16 matter (80): 0.16 mistake (81): 0.16 true (82): 0.16 security (83): 0.16 dole (84): 0.16 hostages (85): 0.16 possibly (86): 0.16 democratic (87): 0.16 dismissed (88): 0.16 john (89): 0.16 criminal (90): 0.16 voted (91): 0.16 grant (92): 0.16 investigations (93): 0.15 knowledge (94): 0.15 repeatedly (95): 0.15 fighting (96): 0.15 robert (97): 0.15 assistant (98): 0.15 deputy (99): 0.15 responsibility (100): 0.15 $\rangle$

The information flow model discovers/boosts in rankings the relevant information contained by the Iran-contra scandal: U.S. government and President Reagan were involved in the illegal arms sales to Iran during the Iran-Iraq war; Reagan was investigated by the congress (Tower commission); profits from the illegal sales were diverted to a secret account to support Nicaraguan rebels; Poindexter, Reagan's National Security Advisor, was involved and resigned with immunity; and so on.

5.5 Inference via Singular Value Decomposition (SVD)

The basic idea motivating the use of singular value decomposition in latent semantic analysis is that the term-document space manifest in document collections is produced by some analogue to a lower dimensional space existing in the minds of the authors and readers of documents in that collection [4]. This motivating assumption has been explored in a series of psycholinguistic experiments conducted over the past couple of decades.

In latent semantic analysis a term-document matrix, A is constructed, in which the matrix elements (a_{td}) express the strength of association between term t and document d. Commonly this measure is derived from raw term-frequencies or weighted term-frequencies measures such as tf$*$idf. Dimensionality reduction is performed using the singular value decomposition.

$$A = U\Sigma V^T \tag{9}$$

where Σ is diagonal with monotonically increasing diagonal values, and U and V are orthogonal. It can be shown that the closest k-rank approximation (in the sense of the matrix 2-norm) to A is achieved by

$$A_k = U_k\Sigma_k V_k^T \tag{10}$$

where U_k and V_k are constructed from the first k columns of U and V respectively and Σ_k is constructed from the first k rows and columns of Σ.

By constructing a low dimensional approximation to A, namely A_k, LSA draws out correlations existing in A and applies them to documents and terms thus creating non-zero entries where previously there were none. For example if it is often the case that "reagan" occurs with "trade" in A then this correlation will be encoded in the columns of U. Then, when during dimensionality reduction, a document mentioning "reagan" but not mentioning "trade" is projected onto these columns of U, it will pickup entries for "trade". Similarly anti-correlations are also induced. If "reagan" and "trade" do not co-occur often in the documents then these words will occur with opposite sign in the columns of U and a document mentioning "reagan" may have its association with "trade" reduced.

Although LSA has traditionally been applied to term-document matrices it is also possible to apply the technique to draw correlations from the HAL matrix. To see the action of LSA in drawing out correlations consider the following artificial example:

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	1	1	1	1	1	1
t_2	1	1	1	1	0	1
t_3	1	1	1	1	1	0
t_4	1	0	1	0	1	0
t_5	0	1	0	1	0	1

After performing singular value decomposition the first three columns of U are

$$\begin{aligned}
\mathbf{u_1} &= \langle -0.57, -0.50, -0.50, -0.29, -0.29 \rangle \\
\mathbf{u_2} &= \langle 0, 0.27, -0.27, -0.65, 0.65 \rangle \\
\mathbf{u_3} &= \langle 0, 0.65, -0.65, 0.27, -0.27 \rangle
\end{aligned}$$

The first column essentially encodes a correlation between all rows of A, saying that if the term corresponding to any row is present, then that is information that each of the other rows will be present to a greater or lesser degree.

The second column essentially encodes the opposition of the fourth and fifth rows which are mutually exclusive in the input data. It also encodes opposition between the second and third rows and correlations between the third and fourth rows and the second and fifth rows. The third vector again encodes opposition between the second and third and forth and fifth but the relation between the second and fourth and third and fifth rows has swapped from opposition to correlation. This swapping of correlation vs. opposition can be seen as corrections for previous decisions.

Dimensionality reduction selects the first k-vectors and so enforces their correlations and oppositions. The high dimensional vectors that have been thrown away encode corrections to these earlier correlations. The 1, 2, and 3-dimensional approximations to the 5'th column of A are shown below together with the 5'th column of A.

$$\begin{aligned}
\mathbf{a_5} &= \langle 1, 0, 1, 1, 0 \rangle \\
\mathbf{a_5^1} &= \langle 0.79, 0.68, 0.68, 0.39, 0.39 \rangle \\
\mathbf{a_5^2} &= \langle 0.79, 0.43, 0.93, 0.99, -0.20 \rangle \\
\mathbf{a_5^3} &= \langle 0.79, 0.18, 1.18, 0.89, -0.10 \rangle
\end{aligned}$$

The 1-dimensional approximation of the 5'th column of A draws a high value for the second row (see the second entry in $\mathbf{a}_5^1$) which wasn't in the initial data. This is due to the high correlation between the first three rows. As the number of dimensions increases the value for this second entry goes down as the higher order vectors account for the error in the assumption of a correlation between the first three rows.

The major difficulty of LSA is the choice of a suitable value for k. A very high k will result in a matrix, A_k very close to the original matrix. A low k will result in a large deviation. Approaches to the choice of k can be divided into theoretical, for example the work by Ding [3], and experimental where an optimal k is derived by reference to some experiment, for example TOEFL vocabulary test reported by Landauer et al [4].

LSA can be used to impose correlations drawn from the HAL matrix to derive a new HAL matrix in a manner similar that of applying weightings to the elements of the HAL matrix. LSA however is quite different in its character to either IVF or TFIDF and in fact is generally combined with TFIDF when applied to term-document matrices.

6 Discussion: Interactively Discovering the Retrieval Context

From a more general perspective, the combination of different types of inference described in the previous section may be applied to discover the retrieval context interactively from some given starting point (i.e., the initial query). Figure 1 illustrates how it goes from "Reagan".

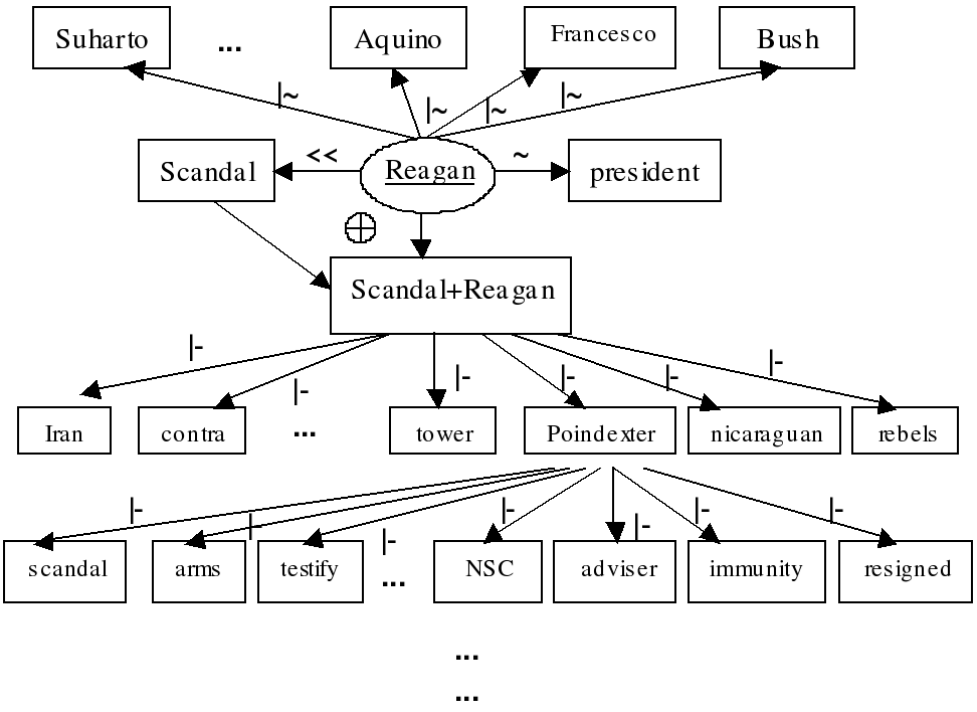

Figure 1: **A Demonstration of discovering retrieval context from "Reagan"**

7 Conclusions and Future Work

This paper addresses a challenging problem encountered in current logic based information retrieval — the automatic construction and maintenance of the background knowledge (termed as information inference in general). To tackle this problem, we have investigated various information inference mechanisms based on a high dimensional semantic space constructed by the Hyperspace Analogue to Language model. An Inverse Vector Frequency method is suggested for smoothing the high frequency bias inherent in the HAL model. The inference mechanisms addressed in this paper are classified as: lexical association, information containment, or information overlapping based. Additionally, the SVD algorithm provides an alternative way to enhance the quality of the HAL matrix as well as infering implicit associations. The different characteristics of these inference mechanisms are demonstrated using examples from the Reuters collection. However, the effectiveness of using one or more of the proposed mechanisms to facilitate the information transformation in logical IR models still needs to be experimentally evaluated on large collections. This is left as future work.

Acknowledgements

The work reported in this paper has been funded in part by the Co-operative Centre for Enterprise Distributed Systems Technology (DSTC) through the Australian Federal Government's CRC Programme (Department of Education, Science and Training).

References

[1] P.D. Bruza and D. Song. Inferring Query Models by Computing Information Flow. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM 2002)*, pages 260–269. ACM Press, 2002.

[2] C. Burgess, K. Livesay, and K. Lund. Explorations in context space: words, sentences, discourse. *Discourse Processes*, 25(2&3):211–257, 1998.

[3] C. H. Q. Ding. A similarity-based probability model for latent semantic indexing. In *SIGIR-1999*, pages 58–65. ACM Press, 1999.

[4] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[5] R. Lau, P. D. Bruza, and D. Song. Belief revision for adaptive information retrieval. In *Proceedings of ACM/SIGIR'04*, 2004.

[6] W. Lowe. Towards a theory of semantic space. In J. D. Moore and K. Stenning, editors, *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, pages 576–581. Lawrence Erlbaum Associates, 2001.

[7] D.W. Song and P.D. Bruza. Discovering Information Flow using a High Dimensional Conceptual Space. In *Proceedings of the 24th Annual ACM Conference of Research and Development in Information Retrieval (SIGIR'2001)*, pages 327–333. ACM Press, 2001.

[8] D.W. Song and P.D. Bruza. Towards context sensitive information inference. *Journal of the American Society for Information Science and Tecnology*, 54(3):321–334, 2003.

[9] C. J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29(6):481–485, 1986.