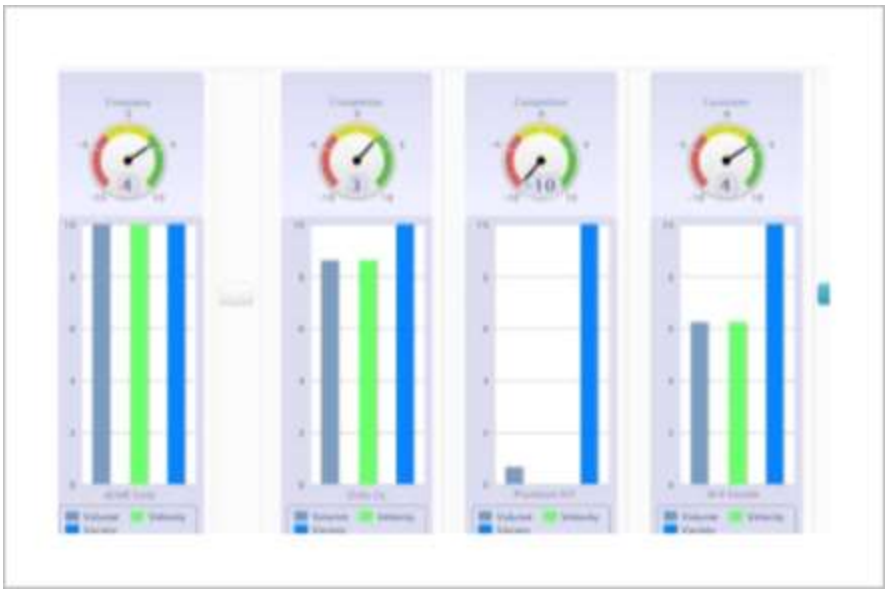

Trend™ Software as a Service

QUICK ARCHITECTURE GUIDE

Robert A Downing II | Copyright | October 14, 2013

Introduction

Trend™ is a Software-as-a-Service [SaaS] family of offerings from the Trendicy Corporation that applies our proprietary technology to the analysis of structured and unstructured data. This analysis can and will yield existing or developing *Trend*s that can be used to monitor the state of the target data items over time, producing *Trend* information that exposes changes in the negative or positive direction of those items.

The Engines

Trend™ is powered by three main engines, which implement the intellectual property that is (or has) been submitted by Trendicy to the US Patent Office. These three engines are fundamental in their design, allowing them to be packaged for invocation and execution by any of the *Trend*™ applications.

THE 'BOT' ENGINE

As discussed above, a great deal of Big Data resides in the Internet, in either Public or Private Domains. The Bot engine is responsible for:

1. Extracting the user-supplied information that is then assembled into arguments to be presented to the various search engines and APIs used by the system;
2. Parsing the linked indices from the search engines to extract the sources [URLs] for the information found;
3. Using the link information to retrieve the source data, or;
4. Using the configured 'special-interest' URLs to retrieve source data;
5. Putting the retrieved information in the appropriate database structures.

THE ANALYSIS ENGINE

The analysis engine is the primary repository for the IP of the company. It is this engine that:

1. Uses the 'found' URLs to solicit the Internet page_ranks associated with each of the URLs. It is this page_rank information that is used in the Impact calculation, allowing the system to assign an impact to each mention found. ***This architectural assumption that varying sites contribute (or detract from) the impact of an item of information is unique to* Trendicy*, and is a major differentiator in the market;***
2. Parses the information found to extract the text that immediately surrounds the search criteria. The assumption here is that a given number of bytes immediately preceding or following a keyword (or keywords) are intimately associated with that word, yielding a 'sentiment score' that conveys the non-textual meaning behind the keyword(s). ***This quantification of the information, in conjunction with the impact, defines the urgency of the information, a measure unique to* Trendicy*, and is a major differentiator in the market;***
3. Establishes the common themes among mentions, linking them according to those themes, such that any mention belongs to at least one theme group, or 'topic,' so that a cumulative sentiment value and impact measure can be applied to the entire topic, ***defining the***

urgency of the information, a measure unique to* Trendicy*, and a major differentiator in the market;

4. Builds the 'local lexicon' which is responsible for implementing the user-context for the evaluation of mentions. While the great majority of a language's vocabulary is constant in its meaning across individuals and populations, there are differences unique to the context in which the vocabulary is used. The context established by the user [see Tools] must be taken into account in presenting an accurate reflection of the attributes attached to information found on the user's behalf. ***This analysis in the user-context is unique to* Trendicy*, and a major differentiator in the market;***
5. Performs the supporting calculations that underpin the Predictive and Prescriptive Analytics of the system ***in the user-context, unique to* Trendicy*, and a major differentiator in the market***:
    a. Statistical analysis of the data captured or generated;
    b. Discovery and confirmation of correlations underlying the data, and;
    c. Generation and reduction of the data to appropriate *Trend* vectors.

THE PRESENTATION ENGINE

It is the presentation engine that is responsible for the gathering of information from the user, placing that information in the appropriate database structure, and then retrieving results from the database for either user consumption or action. The presentation engine:

1. Provides the forms for user entry of:
    a. identifying information and [at least] initial keywords for gathering and assembling data of interest to the user;
    b. a set of defined relationships between entities that constrain interpretations (i.e.: a rule-set);
2. Stores that information in the appropriate database structures for use by the Bot Engine, the Analysis Engine and for later recall as elements of the results displays;
3. Assembles the results of the Analysis Engine for use in the results displays, and;
4. Implements the Predictive and Prescriptive analytics according to the defined user context:
    a. The rule-set configured by the user (e.g.: a commonly-understood Strengths/Weaknesses/Opportunities/Threats rule-set, or; a cause-and-effect rule-set appropriate to a clinical setting) , and;
    b. The depiction of the results in within the bounds of the user context as defined by the lexical and rule-set values;
5. Implements the logic to administer the local lexicon, allowing for the evolution of dialect as the user context evolves over time;

THE IP

Using the system

DATA

Data exists all around us, whether in private locations (like restricted on-line communities) or in the public domain (that is, the Internet). While it may appear that there is little that can be learned

from all of that, it is not the data that is overwhelming. Rather, it is the sheer volume of data that surrounds us, and causes the user to ask “Where do I begin?”

Often, the beginning is with a static set of data, perhaps downloaded from a site providing structured or unstructured data, perhaps downloaded from a medical repository, or perhaps accumulated from surveys or other sources of unstructured data. This type of initial dataset is advantageous because it allows asking the question “Where do I begin?” with boundaries to the answer. The difficulties imposed by dynamic, rapidly growing or changing datasets are not present.

The best approach is for the user to decide “What do I want to know?” Answering this will, more often than not, determine what type of data is needed; structured (static, dependable answers that have little or no variation) or unstructured (basically, data that appears to be like human speech; highly variable, of any length, and containing words that are typical of the person using them with cultural and other influences).

Further, it will also help define where to look for that data: if it’s structured it probably exists in nice, ordered, predictable fashion in a specific location that changes very infrequently. If the data is unstructured, it may exist in any number of locations, of varying lengths, and be connected (or not) to other surrounding data only by chance.

Finding the data

Finding the data to answer your questions can appear to be an insurmountable task. After all, there’s a lot of it, it’s all over the place, and separating what you want from all the rest is something never done before.

The quick answer is to sit down and begin to ‘search’ the locations where the data exists, and then check the results one-by-one to see if they might provide anything useful;

- Select a Public [Google/Bing] or Private [www.ietf.org] source
- Type in your search argument
- Go through the results one-by-one
- Repeat (probably!)

If you’re lucky, they do and you can go on to the more important task(s) of using that data to provide value. However, odds are that you’ll have to do a number of searches to find your answers. Given that the saying “time is money” is as true now as it was when first uttered, this can be an expensive proposition.

The answer is a mechanism or application that can be given the task of automating these repetitive tasks for you, locating, importing and then analyzing the data so that the decision about the value of that data becomes a simple one, dependable and repeatable.

Using the data

Having found interesting data, then, what do you do with it? The goal would be to present the data in a series of graphical displays, which visualize the underlying data in a way designed to enforce the meaning discovered/extracted from the data.

If a “picture is worth a thousand words” then displaying the results in as intuitive a fashion will probably achieve the best results.

INITIALIZING THE SYSTEM

It's necessary to have some very basic information to get *Trend*™ to work. First, you need to give the application something to look for, just as you would have with any available search engine. Prior to your having signed in, some work has been done to identify the 'things' (we'll refer to them as entities) that are of interest to the administrators of your structured and unstructured.

What this means is that you have data to start with, and are ahead of the game. Your displays are already populated and giving you the foundation for constructing the queries you need to extract all the possible meaning you can from your data.

GETTING RESULTS

Your results are presented to you from a bunch of different perspectives, and on a number of different pages. Taken singly, they are significant, but in aggregate you will be able to find a lot of meaning underlying the source data.

Where are they located?

The results are shown in a number of locations:

- The dashboards contain graphical representations that demonstrate results on two different scales:
    - There are several dashboards:
    - The dial-type graphic depicts the overall 'health' of each category [a product of the cumulative measures of sentiment and impact as found in each mention]
    - A vertical bar-type graphic depicts the measures of the data related to each category in terms of three common metrics:
        - Mentions – the total number of mentions about the entity
        - Frequency – the likelihood of something being said about the entity
        - Topics – the 'groupings' into which individual data items can be placed
- The tools have several functions that display results:
    - The Statistics tab shows the basic relationships that have been calculated for each category using standard statistical algorithms
    - The Lexicon displays all the words that were in the 'default' lexicon [dictionary] as well as all the [new] words that have been found in the underlying membership data on three axes:
        - The 'Count' field contains the absolute number of times that word occurred in all of the membership data
        - The 'sent' field contains the *initial* sentiment (i.e.: a negative/neutral/positive score that reflects usage in the general population)
        - The 'LexCat' field contains a number that designates the function of that word (e.g.: it is a neutral, scored word, or it is a diminisher that lessens the impact of an associated word; it is one of: scored word; diminisher; intensifier; negator; new word, or; ignored word)
    - The Links tab would display the URLs from which data was accumulated in the case that the source data came from the Internet. [In the case of a closed community, the source data might be a static file extracted from the community site, so there would be no URLs to display.]

What are they telling me?

They are telling you in a graphical format how each category is perceived by the membership, and also how they compare to each other. The comparison of the graphs to each other can show areas of high interest, areas of low interest, and initial correlations between categories.

How do I use them?

Each set of results can be used individually, to get a single-perspective look at how the membership feel about that particular item. However, the greater strength will lie in using them in conjunction, to get a feeling for the interrelationships that exist deeper in the data.

TOOLS

The tools deserve a special treatment because they give visibility into the underlying relationships in the data (i.e.: the statistics), how the results are grouped and displayed (i.e.: customization), and how to control how the calculations are done (i.e.: the Lexicon and the Links).

Statistics

The statistics shown are those that apply to the significant metrics in the system (e.g.: numbers off mentions found, or the number for each entity; the averages for each metric; correlations between metrics). They are fairly standard statistics, such as 'Arithmetic Mean' [average], 'Standard Deviation,' 'Variance' or 'Correlation.'

- The arithmetic mean depicts the 'average' sentiment associated with each category (e.g.: word count, or sentiment). You'll see additional means being calculated, and they all have a value, but are more specialized in their usage.
- The deviations depict the lower- and upper-bounds around the mean, within which should like 90% of the responses for that category. For example, a mean of negative ninety-eight hundredths [-0.98] with a deviation of four [4] means that the bulk of the sentiment discovered in member data lies between a negative four and ninety-eight hundredths [-4.98] and positive three and two hundredths [+3.02]. The interpretation here is that the overwhelming sentiment is pretty neutral.
- The variances show whether the distribution of the data is fairly smooth and regular (i.e.: a standard 'bell-curve' or Poisson distribution) or there is a long 'tail' to the negative or the positive values. Using the above example, the variance is fifteen and ninety-nine hundredths [15.99] showing that of the outlying values, they tend toward the more positive than negative.
- 'Correlations' depict the relative strengths of relationships between data. A negative correlation means that there is an inverse relationship between the data to some degree. Likewise, a positive correlation means that there is a direct relationship between the data. In the example above, the correlation between 'Sentiment & Lexical value' shows that there exists a positive relationship between the sentiments of the words that are contained in the member data, and the actual type of word: in this case, the more commonly occurring words are modifiers (either diminishers or magnifiers).

Customize

The customization function allows you to change the labels being used in the application. Changes here will not affect the underlying data, but can be used to make the representations of that data more intuitive. These are text-entry fields, and if changed, must be “Updated” to take effect.

Lexicon

As discussed, the Lexicon gives you visibility into the language used to convey the characteristics of the membership. The starting point for the Lexicon was a dictionary of commonly-occurring words with their associated sentiment values and parts of speech [as established by the Massachusetts Institute of Technology].

While these values are typical of an English-speaking population covering all demographics, they are not necessarily accurate for a closed population having one or more traits in common. Therefore, the Lexicon allows for modification of words to make them more contextual to the underlying data, and make the resulting calculations more accurate. Both sentiment value and lexical category can be changed to establish a greater relevance to the structured and unstructured.

Clicking on a word brings up the dialogs necessary to change these values. Sentiment values have a range between a negative ten [-10] and positive ten [10]. Lexicon category has a range of zero [0] to six [6]. It is recommended that care be taken in making changes in this area as not only will it have a direct result on the calculations, but will also potentially skew the results beyond what is intended.

NOTE: At this time, changes to the lexicon will not be effective until a recalculation is requested. While the long-term intent is to provide for the automation of this request, for the present a rescoring of the database must be requested of a Trendicy representative.

Links

The links represent the sources where we find our (your) information. They are most commonly URLs, but may be pointers to other types of sources, like a user's "news feed" from Facebook, or specific conversation from a community board. In any event, they are items that can be selected (or deselected) according to their perceived value on the part of the user.

Using the system

DATA

Data exists all around us, whether in private locations (like restricted on-line communities) or in the public domain (that is, the Internet). While it may appear that there is little that can be learned from all of that, it is not the data that is overwhelming. Rather, it is the sheer volume of data that surrounds us, and causes the user to ask “Where do I begin?”

Often, the beginning is with a static set of data, perhaps downloaded from a site providing structured or unstructured data, perhaps downloaded from a medical repository, or perhaps accumulated from surveys or other sources of unstructured data. This type of initial dataset is advantageous because it allows asking the question “Where do I begin?” with boundaries to the answer. The difficulties imposed by dynamic, rapidly growing or changing datasets are not present.

The best approach is for the user to decide "What do I want to know?" Answering this will, more often than not, determine what type of data is needed; structured (static, dependable answers that have little or no variation) or unstructured (basically, data that appears to be like human speech; highly variable, of any length, and containing words that are typical of the person using them with cultural and other influences).

Further, it will also help define where to look for that data: if it's structured it probably exists in nice, ordered, predictable fashion in a specific location that changes very infrequently. If the data is unstructured, it may exist in any number of locations, of varying lengths, and be connected (or not) to other surrounding data only by chance.

Finding the data

Finding the data to answer your questions can appear to be an insurmountable task. After all, there's a lot of it, it's all over the place, and separating what you want from all the rest is something never done before.

The quick answer is to sit down and begin to 'search' the locations where the data exists, and then check the results one-by-one to see if they might provide anything useful;

- Select a Public [Google/Bing] or Private [www.ietf.org] source
- Type in your search argument
- Go through the results one-by-one
- Repeat (probably!)

If you're lucky, they do and you can go on to the more important task(s) of using that data to provide value. However, odds are that you'll have to do a number of searches to find your answers. Given that the saying "time is money" is as true now as it was when first uttered, this can be an expensive proposition.

The answer is a mechanism or application that can be given the task of automating these repetitive tasks for you, locating, importing and then analyzing the data so that the decision about the value of that data becomes a simple one, dependable and repeatable.

Using the data

Having found interesting data, then, what do you do with it? The goal would be to present the data in a series of graphical displays, which visualize the underlying data in a way designed to enforce the meaning discovered/extracted from the data.

If a "picture is worth a thousand words" then displaying the results in as intuitive a fashion will probably achieve the best results.

INITIALIZING THE SYSTEM

It's necessary to have some very basic information to get *Trend*™ to work. First, you need to give the application something to look for, just as you would have with any available search engine. Prior to your having signed in, some work has been done to identify the 'things' (we'll refer to them as entities) that are of interest to the administrators of your structured and unstructured.

What this means is that you have data to start with, and are ahead of the game. Your displays are already populated and giving you the foundation for constructing the queries you need to extract all the possible meaning you can from your data.

GETTING RESULTS

Your results are presented to you from a bunch of different perspectives, and on a number of different pages. Taken singly, they are significant, but in aggregate you will be able to find a lot of meaning underlying the source data.

Where are they located?

The results are shown in a number of locations:

- The dashboards contain graphical representations that demonstrate results on two different scales:
    - There are several dashboards:
    - The dial-type graphic depicts the overall 'health' of each category [a product of the cumulative measures of sentiment and impact as found in each mention]
    - A vertical bar-type graphic depicts the measures of the data related to each category in terms of three common metrics:
        - Mentions – the total number of mentions about the entity
        - Frequency – the likelihood of something being said about the entity
        - Topics – the 'groupings' into which individual data items can be placed
- The tools have several functions that display results:
    - The Statistics tab shows the basic relationships that have been calculated for each category using standard statistical algorithms
    - The Lexicon displays all the words that were in the 'default' lexicon [dictionary] as well as all the [new] words that have been found in the underlying membership data on three axes:
        - The 'Count' field contains the absolute number of times that word occurred in all of the membership data
        - The 'sent' field contains the *initial* sentiment (i.e.: a negative/neutral/positive score that reflects usage in the general population)
        - The 'LexCat' field contains a number that designates the function of that word (e.g.: it is a neutral, scored word, or it is a diminisher that lessens the impact of an associated word; it is one of: scored word; diminisher; intensifier; negator; new word, or; ignored word)
    - The Links tab would display the URLs from which data was accumulated in the case that the source data came from the Internet. [In the case of a closed community, the source data might be a static file extracted from the community site, so there would be no URLs to display.]

What are they telling me?

They are telling you in a graphical format how each category is perceived by the membership, and also how they compare to each other. The comparison of the graphs to each other can show areas of high interest, areas of low interest, and initial correlations between categories.

How do I use them?

Each set of results can be used individually, to get a single-perspective look at how the membership feel about that particular item. However, the greater strength will lie in using them in conjunction, to get a feeling for the interrelationships that exist deeper in the data.

TOOLS

The tools deserve a special treatment because they give visibility into the underlying relationships in the data (i.e.: the statistics), how the results are grouped and displayed (i.e.: customization), and how to control how the calculations are done (i.e.: the Lexicon and the Links).

Statistics

The statistics shown are those that apply to the significant metrics in the system (e.g.: numbers off mentions found, or the number for each entity; the averages for each metric; correlations between metrics). They are fairly standard statistics, such as 'Arithmetic Mean' [average], 'Standard Deviation,' 'Variance' or 'Correlation.'

- The arithmetic mean depicts the 'average' sentiment associated with each category (e.g.: word count, or sentiment). You'll see additional means being calculated, and they all have a value, but are more specialized in their usage.
- The deviations depict the lower- and upper-bounds around the mean, within which should like 90% of the responses for that category. For example, a mean of negative ninety-eight hundredths [-0.98] with a deviation of four [4] means that the bulk of the sentiment discovered in member data lies between a negative four and ninety-eight hundredths [-4.98] and positive three and two hundredths [+3.02]. The interpretation here is that the overwhelming sentiment is pretty neutral.
- The variances show whether the distribution of the data is fairly smooth and regular (i.e.: a standard 'bell-curve' or Poisson distribution) or there is a long 'tail' to the negative or the positive values. Using the above example, the variance is fifteen and ninety-nine hundredths [15.99] showing that of the outlying values, they tend toward the more positive than negative.
- 'Correlations' depict the relative strengths of relationships between data. A negative correlation means that there is an inverse relationship between the data to some degree. Likewise, a positive correlation means that there is a direct relationship between the data. In the example above, the correlation between 'Sentiment & Lexical value' shows that there exists a positive relationship between the sentiments of the words that are contained in the member data, and the actual type of word: in this case, the more commonly occurring words are modifiers (either diminishers or magnifiers).

Customize

The customization function allows you to change the labels being used in the application. Changes here will not affect the underlying data, but can be used to make the representations of that data more intuitive. These are text-entry fields, and if changed, must be "Updated" to take effect.

Lexicon

As discussed, the Lexicon gives you visibility into the language used to convey the characteristics of the membership. The starting point for the Lexicon was a dictionary of commonly-occurring words

with their associated sentiment values and parts of speech [as established by the Massachusetts Institute of Technology].

While these values are typical of an English-speaking population covering all demographics, they are not necessarily accurate for a closed population having one or more traits in common. Therefore, the Lexicon allows for modification of words to make them more contextual to the underlying data, and make the resulting calculations more accurate. Both sentiment value and lexical category can be changed to establish a greater relevance to the structured and unstructured.

Clicking on a word brings up the dialogs necessary to change these values. Sentiment values have a range between a negative ten [-10] and positive ten [10]. Lexicon category has a range of zero [0] to six [6]. It is recommended that care be taken in making changes in this area as not only will it have a direct result on the calculations, but will also potentially skew the results beyond what is intended.

NOTE: At this time, changes to the lexicon will not be effective until a recalculation is requested. While the long-term intent is to provide for the automation of this request, for the present a rescoring of the database must be requested of a Trendicy representative.

Links

The links represent the sources where we find our (your) information. They are most commonly URLs, but may be pointers to other types of sources, like a user's "news feed" from Facebook, or specific conversation from a community board. In any event, they are items that can be selected (or deselected) according to their perceived value on the part of the user.